

Complexity

Emerging Applications of Complex Networks

Lead Guest Editor: Gerard Olivar-Tost

Guest Editors: Jesús Gómez-Gardeñes and Rafael Hurtado-Heredia





Emerging Applications of Complex Networks

Complexity

Emerging Applications of Complex Networks

Lead Guest Editor: Gerard Olivar-Tost

Guest Editors: Jesús Gómez-Gardeñes and Rafael Hurtado-Heredia



Copyright © 2018 Hindawi. All rights reserved.


This is a special issue published in “Complexity.” All articles are open access articles distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Editorial Board



- José A. Acosta, Spain
Carlos F. Aguilar-Ibáñez, Mexico
Mojtaba Ahmadih Khanesar, UK
Tarek Ahmed-Ali, France
Alex Alexandridis, Greece
Basil M. Al-Hadithi, Spain
Juan A. Almendral, Spain
Diego R. Amancio, Brazil
David Arroyo, Spain
Mohamed Boutayeb, France
Átila Bueno, Brazil
Arturo Buscarino, Italy
Guido Caldarelli, Italy
Eric Campos-Canton, Mexico
Mohammed Chadli, France
Émile J. L. Chappin, Netherlands
Diyi Chen, China
Yu-Wang Chen, UK
Giulio Cimini, Italy
Danilo Comminiello, Italy
Sara Dadras, USA
Sergey Dashkovskiy, Germany
Manlio De Domenico, Italy
Pietro De Lellis, Italy
Albert Diaz-Guilera, Spain
Thach Ngoc Dinh, France
Jordi Duch, Spain
Marcio Eisencraft, Brazil
Joshua Epstein, USA
Mondher Farza, France
Thierry Floquet, France
Mattia Frasca, Italy
José Manuel Galán, Spain
Lucia Valentina Gambuzza, Italy
Bernhard C. Geiger, Austria
Carlos Gershenson, Mexico
Peter Giesl, UK
Sergio Gómez, Spain
Lingzhong Guo, UK
Xianggui Guo, China
Sigurdur F. Hafstein, Iceland
Chittaranjan Hens, Israel
Giacomo Innocenti, Italy
Sarangapani Jagannathan, USA
Mahdi Jalili, Australia
Jeffrey H. Johnson, UK
M. Hassan Khooban, Denmark
Abbas Khosravi, Australia
Toshikazu Kuniya, Japan
Vincent Labatut, France
Lucas Lacasa, UK
Guang Li, UK
Qingdu Li, Germany
Chongyang Liu, China
Xiaoping Liu, Canada
Xinzhi Liu, Canada
Rosa M. Lopez Gutierrez, Mexico
Vittorio Loreto, Italy
Noureddine Manamanni, France
Didier Maquin, France
Eulalia Martínez, Spain
Marcelo Messias, Brazil
Ana Meštrović, Croatia
Ludovico Minati, Japan
Ch. P. Monterola, Philippines
Marcin Mrugalski, Poland
Roberto Natella, Italy
Sing Kiong Nguang, New Zealand
Nam-Phong Nguyen, USA
B. M. Ombuki-Berman, Canada
Irene Otero-Muras, Spain
Yongping Pan, Singapore
Daniela Paolotti, Italy
Cornelio Posadas-Castillo, Mexico
Mahardhika Pratama, Singapore
Luis M. Rocha, USA
Miguel Romance, Spain
Avimanyu Sahoo, USA
Matilde Santos, Spain
Josep Sardanyés Cayuela, Spain
Ramaswamy Savitha, Singapore
Hiroki Sayama, USA
Michele Scarpiniti, Italy
Enzo Pasquale Scilingo, Italy
Dan Selişteanu, Romania
Dehua Shen, China
Dimitrios Stamovlasis, Greece
Samuel Stanton, USA
Roberto Tonelli, Italy
Shahadat Uddin, Australia
Gaetano Valenza, Italy
Dimitri Volchenkov, USA
Christos Volos, Greece
Zidong Wang, UK
Yan-Ling Wei, Singapore
Honglei Xu, Australia
Yong Xu, China
Xinggang Yan, UK
Baris Yuçe, UK
Massimiliano Zanin, Spain
Hassan Zargarzadeh, USA
Rongqing Zhang, USA
Xianming Zhang, Australia
Xiaopeng Zhao, USA
Quanmin Zhu, UK

Contents

Emerging Applications of Complex Networks

Gerard Olivar-Tost , Jesús Gómez-Gardeñes, and Rafael Hurtado-Heredia
Editorial (2 pages), Article ID 8513082, Volume 2018 (2018)


A Network-Based Approach to Modeling and Predicting Product Coconsideration Relations

Zhenghui Sha, Yun Huang , Jiawei Sophia Fu, Mingxian Wang, Yan Fu, Noshir Contractor,
and Wei Chen 
Research Article (14 pages), Article ID 2753638, Volume 2018 (2018)

Bipartisanship Breakdown, Functional Networks, and Forensic Analysis in Spanish 2015 and 2016 National Elections

Juan Fernández-Gracia and Lucas Lacasa 
Research Article (23 pages), Article ID 9684749, Volume 2018 (2018)

Spatial “Artistic” Networks: From Deconstructing Integer-Functions to Visual Arts

Ernesto Estrada  and Puri Pereira-Ramos
Research Article (8 pages), Article ID 9893867, Volume 2018 (2018)

Modeling and Simulation of Project Management through the PMBOK® Standard Using Complex Networks

Luz Stella Cardona-Meza and Gerard Olivar-Tost
Research Article (12 pages), Article ID 4791635, Volume 2017 (2018)

A Language as a Self-Organized Critical System

Vasilii A. Gromov and Anastasia M. Migrina
Research Article (7 pages), Article ID 9212538, Volume 2017 (2018)

An Approach for Understanding and Promoting Coal Mine Safety by Exploring Coal Mine Risk Network

Yongliang Deng, Liangliang Song, Zhipeng Zhou, and Ping Liu
Research Article (17 pages), Article ID 7628569, Volume 2017 (2018)

Editorial

Emerging Applications of Complex Networks

Gerard Olivar-Tost ¹, **Jesús Gómez-Gardeñes**,² and **Rafael Hurtado-Heredia**³

¹*Department of Mathematics, Faculty of Sciences, Universidad Nacional de Colombia, Manizales, Colombia*

²*Institute for Biocomputation & Physics of Complex Systems, Faculty of Sciences, University of Zaragoza, Zaragoza, Spain*

³*Department of Physics, Faculty of Sciences, Universidad Nacional de Colombia, Bogotá, Colombia*

Correspondence should be addressed to Gerard Olivar-Tost; golivart@unal.edu.co

Received 28 November 2018; Accepted 28 November 2018; Published 11 December 2018

Copyright © 2018 Gerard Olivar-Tost et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A large amount of research is available on complex networks modeling and simulation, with applications to smart grids and energy, vehicular traffic, diseases and epidemics, ecosystems, supply chains, financial systems, social science, urban environment, and so on. However, there are still some topics where complex networks are just starting to pursue their role.

With new hybrid mathematical modeling tools and the power of computation, there is a broad range of emerging applications. Thus novel research focuses on these topics, like arts and literature, human senses, medicine and physiology, psychology, lifestyle, philosophy, and project management.

This special issue holds 6 original research articles and various applications regarding products, voting, arts, project management, language, and safety as emergent applications of Complex Networks.

Z. Sha et al. showed how understanding customer preferences in consideration decisions is critical to choice modeling in engineering design. In their paper, they presented a network-based approach based on Exponential Random Graph Models to study customers' consideration behaviors according to engineering design. Their approach is capable of modeling the endogenous effects among products through various network structures (e.g., stars and triangles) besides the exogenous effects and predicting whether two products would be considered together. Using buyer survey data from the China automarket in 2013 and 2014, they evaluate the goodness of fit and the predictive power of the two models.

J. Fernández-Gracia and L. Lacasa present a social network and forensic analysis of the vote counts of Spanish national elections that took place in December 2015 and their

sequel in June 2016. They initially consider the phenomenon of bipartisanship breakdown and find that such breakdown is more prominently close to cosmopolite and largely populated areas and less important in rural areas. Through functional network analysis they detect an effective partition of municipalities which remarkably coincides with the first-level political and administrative division of autonomous communities. Finally, they further explore the cooccurring statistics of vote share and turnout, finding a mild tendency in the clusters of the conservative party to smear out towards the area of high turnout and vote share, what has been previously interpreted as a possible sign of incremental fraud.

E. Estrada and P. Pereira-Ramos produced a relevant and very interesting and original document on deconstructivism, as an aesthetically appealing architectonic style. They identify some general characteristics of this style, such as decomposition of the whole into parts, superposition of layers, and conservation of the memory of the whole. Using these attributes, they propose a method to deconstruct functions based on integers and generate spatial networks which display a few artistic attributes such as (i) biomorphic shapes, (ii) symmetry, and (iii) beauty. They show how these networks inspire an artist to create artistic compositions using mixed techniques on canvas and on paper. They specially claim that the aesthetic of network research, and not only its applicability, would be an attractor for new minds to this field.

L. S. Cardona-Meza and G. Olivar-Tost discuss about project management, which is predominantly based on theories of control. In complex environments, management problems arise from assuming that results, predicted at the

start of a project, can be sufficiently described and delivered as planned. Thus, once a project reaches a critical size, a calendar, and a certain level of ambiguity and interconnection, the analysis centered on control does not function adequately. In their study, through a complex network, the dynamic structure of a project and its trajectories are simulated using inference processes. Finally, some numerical simulations are described, leading to a decision making tool that identifies critical processes, thereby obtaining better performance outcomes of projects.

V. A. Gromov and A. M. Migrina produced a paper on natural language (represented by texts generated by native speakers). It is considered as a complex system, and the type thereof to which natural languages belong is ascertained. The authors hypothesize that a language is a self-organized critical system and that the texts of a language are “avalanches” flowing down its word cooccurrence graph. The respective statistical characteristics for distributions of the number of words in the texts of English and Russian languages are calculated. The analysis found that the number of words in the texts obeys power-law distribution.

Y. Deng et al. showed that capturing the interrelations among risks is essential to thoroughly understand and promote coal mining safety. Several parameters were employed to reveal the topological properties of the network. As indicated by the results, the considered network possesses scale-free network property because its cumulative degree distribution obeys power-law distribution. This means that it is robust to random hazard and vulnerable to deliberate attack. Also, it is a small-world network due to its relatively small average path length as well as high clustering coefficient, implying that accident propagation in this type of network is faster than in regular ones. Furthermore, the effect of risk control is explored. According to the result, it shows that roof collapse, fire, and gas concentration exceeding limit refer to three most valuable targets for risk control among all the risks.

In summary, this special issue will provide a nice panorama of the present status of emerging applications and recent developments, giving novel and important insights of management, physical, and art applications.

Conflicts of Interest

The editors declare that there are no conflicts of interest regarding the publication of this editorial.

Acknowledgments

The editors would like to thank all authors who submitted their research to this special issue, as well as all reviewers for their valuable contribution.

*Gerard Olivar-Tost
Jesús Gómez-Gardeñes
Rafael Hurtado-Heredia*

Research Article

A Network-Based Approach to Modeling and Predicting Product Coconsideration Relations

Zhengkui Sha,¹ Yun Huang ,² Jiawei Sophia Fu,³ Mingxian Wang,⁴
Yan Fu,⁴ Noshir Contractor,² and Wei Chen ⁵

¹Department of Mechanical Engineering, University of Arkansas, Fayetteville, AR, USA

²Department of Industrial Engineering & Management Sciences and Department of Management & Organizations and Department of Communication Studies, Northwestern University, Evanston, IL, USA

³Media, Technology, and Society, Northwestern University, Evanston, IL, USA

⁴Global Data Insight & Analytics, Ford Motor Company, Dearborn, MI, USA

⁵Department of Mechanical Engineering, Northwestern University, Evanston, IL, USA

Correspondence should be addressed to Wei Chen; weichen@northwestern.edu

Received 23 September 2017; Accepted 18 December 2017; Published 28 January 2018

Academic Editor: Jesús Gómez-Gardeñes

Copyright © 2018 Zhengkui Sha et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Understanding customer preferences in consideration decisions is critical to choice modeling in engineering design. While existing literature has shown that the exogenous effects (e.g., product and customer attributes) are deciding factors in customers' consideration decisions, it is not clear how the endogenous effects (e.g., the intercompetition among products) would influence such decisions. This paper presents a network-based approach based on Exponential Random Graph Models to study customers' consideration behaviors according to engineering design. Our proposed approach is capable of modeling the endogenous effects among products through various network structures (e.g., stars and triangles) besides the exogenous effects and predicting whether two products would be considered together. To assess the proposed model, we compare it against the dyadic network model that only considers exogenous effects. Using buyer survey data from the China automarket in 2013 and 2014, we evaluate the goodness of fit and the predictive power of the two models. The results show that our model has a better fit and predictive accuracy than the dyadic network model. This underscores the importance of the endogenous effects on customers' consideration decisions. The insights gained from this research help explain how endogenous effects interact with exogenous effects in affecting customers' decision-making.

1. Introduction

Complex network modeling and simulation have shown their power in many engineering applications, such as the wireless network, sensor network, smart grids, supply chain, transportation systems, and many others. Recent developments in mathematical modeling techniques and computational algorithms to study complex networks have also drawn the attention of engineering design field. Complex networks have been used in engineering design for the study of relational patterns, effective network visualization of associations of products, and modeling social interactions [1] and cross-level interactions between customers and products [2, 3]. In the design of complex products, network analysis has been used

to characterize a product as a network of components that share technical interfaces or connections. Various network metrics, such as clustering coefficients and path length, are used to characterize the product structure and study the correlations between design quality and the product structure. Based on the network metrics, for example, the centrality, Sosa et al. [4] defined three measures of modularity as a way to improve the understanding of product architecture. Recent work by Sosa et al. [5] found that proactively managing the use of network structure (such as hubs) may help improve the quality of complex product designs. Network analysis has also been applied to studying designers' network for understanding organizational behavior [6, 7] and improving multidisciplinary design efficiency [8]. In this paper, instead

of focusing on the product or the designer, we leverage complex network modeling and simulation techniques to study another key stakeholder in product design, the customer. We aim to leverage complex networks to study customer preference in support of product design and development. Particularly, in this paper, we study customers' *consideration* decisions by modeling *product coconsideration relations*, two products being concurrently considered in purchase, as a complex network.

2. Background and Literature Review

Choice modeling is of great interest in engineering design as it predicts product demand and market share as a function of engineering design attributes and customer profiles in a target market [9]. Choice models have been integrated into design optimization to take account of customer preferences in supporting engineering design decisions [9–12]. Previous choice models mostly assume that customers have bounded rationality and have underlying utilities to rank alternatives in a *consideration set*, “a set of product alternatives available to an individual who will seriously evaluate through comparisons before making a final choice” [13]. A key step of constructing choice models is to determine the consideration set [14]. As Hauser et al. [15] indicated “if customers do not consider your product, they can't choose it.”

From an enterprise perspective, understanding customer preferences in consideration is important for identifying crucial product features that customers are willing to pay for. Existing studies [16, 17] also revealed the *consideration set phenomenon*, that is, the size of the consideration set tends to be much smaller (roughly 5-6 brands) than the total number of choices available in a market. As a result, small changes in individuals' consideration sets (either size or options) may significantly transform the landscape of the overall market and reshape the competition relations in an existing market. Therefore, understanding customers' preferences in consideration poses new opportunities to optimize product configurations, address customer needs, establish competitive design strategies, and make strategic moves such as branding and positioning.

Managerial actions have been taken to influence customers' consideration decisions directly, for example, by changing brand accessibility [18] and by controlling usage and awareness [19]. However, quantitative studies on customers' consideration decisions are challenging as consideration is an intermediate construct, not the final choice [15]. The decision context and a large amount of uncertainty alter decision rules. Existing literature primarily focuses on inferring decision-rule heuristics [20–22], such as the cognitive simplicity rule [23], which has been shown to be effective in automobile and web-based purchasing. There are three approaches to uncover consideration decision-rule heuristics [15]. The first approach only utilizes final choices and product features in the consideration set. It adopts a two-stage consider-then-choose decision process and infers model parameters using the Bayesian or maximum likelihood estimation. Typical methods include Bayesian [24], choice-set explosion [25–27], and soft constraints [28]. The second approach

measures consideration through designed experiments *in vitro*, similar to the choice-based conjoint analysis exercise [15]. Then the decision rules that best explain the observed consideration decisions are estimated with Bayesian [29] and machine-learning pattern-matching algorithms [30]. The third approach measures decision rules directly through self-explicated questions [31].

Despite the diversity of research on consideration sets, few studies have focused on understanding the underlying process of generating customer consideration sets. The connection between the formation of consideration sets and the driving factors is not well understood. Particularly, we know little about how the inherent market structure, including both *the interdependence among existing products* and *association among customers*, affects the consideration decisions. To address this research gap, we develop a network-based approach to model customers' consideration behaviors by modeling product coconsideration relations. As shown in Figure 1, the key idea of the proposed network approach is to transform customer consideration sets into a product association network, in which nodes represent products and links represent the coconsideration between two products. As a result, the problem of understanding customer consideration can be addressed by predicting certain network structures as a function of association networks formed by product attributes and customer demographics. It is worth noting that as the link formation is an aggregation of customers' decisions, the links (i.e., the coconsideration relations) imply the competition among products. Therefore, our approach enables us to study customer preference and market structure in an integrated manner. This is different from the studies in choice modeling (e.g., the monomial logit choice model [32]) that focus on establishing models for individuals. It is also worth noting that our study is different from the agent-based models which hypothesize certain individual choice-making rules [33]. Instead, our approach is *data-driven*, which leverages the observed data to drive the establishment of coconsideration models and prediction analysis using the estimated model parameters.

Recently, network approaches have been also extensively used in recommender systems [34–38]. Recommender systems are frequently used to recommend products to customers based on what they searched (considered). From the network representation point of view, our approach is similar to the bipartite projection approach [39] used in the recommender systems research. However, the proposed network approach is distinct from network-based recommender algorithms [37, 38] in two aspects: first, the end goal is different. The recommender algorithms attempt to *predict* future likes and interests by mining data on past user activities. Common methods include the similarity-based methods (e.g., the collaborative filtering [38], content-based analysis [40], and Dirichlet allocation [41]) and the recently developed hybrid methods [36, 42]. The approach proposed in this paper relies on the network-based *statistical inference model*, which emphasizes *deduction* and *explanation*. It aims to provide an explanatory framework for customers' consideration behaviors, so that a feedback loop can be created from customer preference to engineering design. Therefore, the

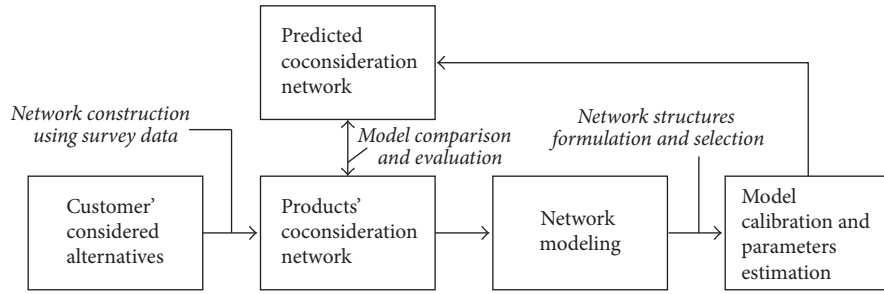


FIGURE 1: The research approach and the research focus.

end goal of this study is to inform product design for larger market share. In such a context, prediction in this study is for comparison and validation purposes. Second, the role of network in the modeling is different. In existing network-based recommender algorithms, the input takes various graph-based node-specific attributes (e.g., degree), which are essentially the exogenous factors, to generate the similarity metrics. In our approach, the model input can take into account present network structures (e.g., triangles and loops), which represents the interdependencies among products, so that the effect of the inherent competition relations can be assessed. Such a capability supports better understanding on the consideration behaviors and could provide additional insights into the design research that has been primarily driven by users' preferences to engineering attributes.

The current work builds upon our previous research efforts. In our recent study, Fu et al. [43] developed a two-stage bipartite network modeling approach to study customer preferences in making choices by decoupling the choice-making process in two stages, the consideration stage and the choice-making stage. Wang et al. [44] utilized a *dyadic network* analysis approach to predict product coconsideration relations based on *exogenous factors*, such as product attributes and customer demographics. By mapping specific technological advancement (e.g., turbocharged techniques) to the change of products attributes, the authors also demonstrated how the model facilitates the forecast of the impact of technological changes on product coconsideration and market competition.

In this paper, we take a further step to investigate the power of complex network modeling in understanding product coconsideration relations by considering both *exogenous factors* and *endogenous factors*, for example, product interdependence and inherent market competition. The core technique is based on the Exponential Random Graph Model (ERGM) [45]. While dyadic network models are convenient to predict the associations between products based on exogenous factors, ERGM incorporates endogenous factors as well as other network interdependencies [46].

The *research objective* of this study is therefore twofold: (a) to establish the network-modeling framework that supports the explanation of customer's consideration behaviors and enables the prediction of future market competitions; (b) to compare the ERGM and dyadic network model to examine if the inclusion of product interdependence through the

endogenous network effects would better capture the dynamics underlying the formation of product coconsideration relations. The remainder of the paper has five sections. Section 3 presents the research problem and introduces the method of constructing a product coconsideration network. We also briefly provide the technical background of the dyadic network model and ERGM. Section 4 describes the vehicle case study and the data source. We present the estimation results of the dyadic model and ERGM and illustrate how to use the attribute-related network structures to represent product interdependence, that is, the endogenous effects. To evaluate the performance of each model, Section 5 assesses model fit at both the global network level and the local link level. Section 6 evaluates the performance of each model in predicting future coconsideration relations. Finally, Section 7 presents practical implications of the findings and directions for future research.

3. Network Construction and Introduction to Network Models

3.1. Network Construction. The product coconsideration network is constructed using data from customers' consideration sets. The presence of a link (i.e., coconsideration) between two nodes (i.e., products) is determined by an association metric, called *lift* [47]. Equation (1) defines the *lift* value between products i and j . Similar to pointwise mutual information [48], *lift* measures the likelihood of the coconsideration of two products given their individual frequencies of considerations.

$$\text{lift}(i, j) = \frac{\Pr(i, j)}{\Pr(i) \cdot \Pr(j)}, \quad (1)$$

where $\Pr(i, j)$ is the probability of a pair of products i and j being coconsidered by customers among all possibilities, calculated based on the collected consideration data; and $\Pr(i)$ is the probability of individual product i being considered. The *lift* value indicates how likely two products are coconsidered by all customers at the aggregate level, normalized by the product popularity in the entire market. We use this probability of coconsideration, different from market share that is directly determined by the total purchases, to capture the competition between products.

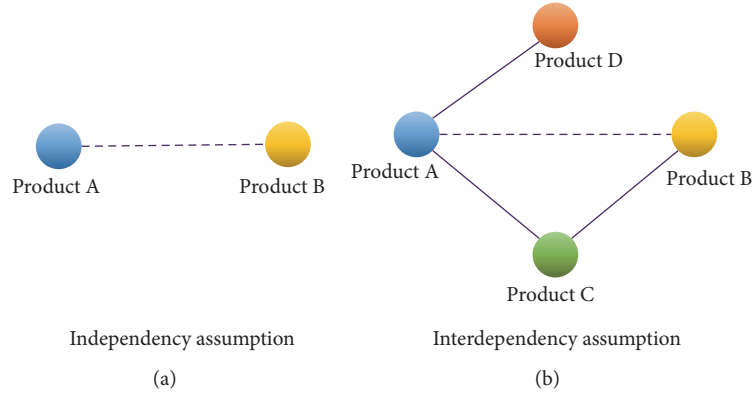


FIGURE 2: Two dependence assumptions underlying the coconsideration network.

With the *lift* value, an undirected coconsideration network can be constructed using the following binary rule:

$$E_{ij} = \begin{cases} 1, & \text{if } \text{lift}(i, j) \geq \text{cutoff} \\ 0, & \text{otherwise,} \end{cases} \quad (2)$$

where *cutoff* is the threshold to determine the presence of a link E_{ij} between two nodes i and j . Statistically, the *lift* value 1 indicates that two products are completely independent [44]; a *lift* value greater than 1 indicates the two products are coconsidered more likely than expected by chance. Based on the application context, research interest, and model requirement, different *lift* values greater than 1 can be used as the cutoff value. Equations (1) and (2) suggest that the network adjacency matrix is symmetric and binary. In this paper, the research is focused on predicting whether two products would have been coconsidered or not. The extent of how often they are coconsidered (reflecting the competition intensity) is not the research focus of this paper. This is why we made the decision of using binary network instead of weighted network. Modeling a binary network, while computationally simpler, is not as rich as the valued network. Hence, we tested the robustness of our findings by estimating multiple models based on varying the cutoff values of *lift*.

3.2. Research Question in the Network Context. Once a coconsideration network is constructed, the likelihood of customers considering two products can be formulated as the probability of a coconsideration link. For prediction purpose, this leads to the question of what factors (e.g., product attributes and customer demographics) drive the formation of a link between a pair of nodes, and how significantly each factor plays a role in the link formation process. The aforementioned *research question* is recast as how to build a network model to predict whether a coconsideration link exists given the network structures, product attributes, and customer profiles.

We posit that there are two decision-making scenarios underlying the coconsideration relations. The first scenario (Figure 2(a)) assumes that each pair of products is independently evaluated by customers. Even for multiple alternatives

in a consideration set, it treats the comparison of each two of these alternatives independent of other pairwise comparisons. The second scenario takes a more general interdependence assumption, where the formation of one coconsideration link is not independent of other coconsideration links. For example, in the right diagram of Figure 2, the likelihood of a coconsideration link between products A and B may be affected by the fact that they are both coconsidered with product C. For the two aforementioned network models, the dyadic network model takes the simple independence assumption, while the ERGM assumes that all coconsideration relations sharing one node are interdependent. In this paper, we will examine whether the ERGM provides a more accurate understanding on the factors driving product coconsiderations by evaluating the goodness of fit and the predictability of the two models.

3.3. Introduction to Network Models. The dyadic network model is analogous to the standard logistic regression element-wise on network matrices, where the model is given by the following:

$$\begin{aligned} \text{logit} [\Pr(Y_{ij} = 1)] &= \boldsymbol{\beta} \mathbf{X}^{(n)} \\ &= \beta_0 + \beta_1 X_{ij}^{(1)} + \dots + \beta_n X_{ij}^{(n)}. \end{aligned} \quad (3)$$

The response Y_{ij} is the binary links E_{ij} between nodes i and j defined in (2). The node attributes are converted to a vector of as *dyadic variable*, $\mathbf{X}^{(n)} = (x_{ij}^{(1)}, \dots, x_{ij}^{(n)})$. Each dyadic variable measures the similarity or difference between pairs of nodes based on the attributes of nodes and a specific arithmetic function (see Table 1 for various dyadic variables). The dyadic network models use the dyadic variables \mathbf{X} to *predict* the complex structures of the observed network composed of coconsideration links. The coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_n)$ indicate the importance of individual dyadic variable in forming a coconsideration relation. Note that, in this model, the probability of each link is evaluated independently.

3.3.1. Exponential Random Graph Model. Other than the dyadic attribute effects, in a network, many links connected

TABLE 1: Constructing explanatory dyadic attributes.

Configuration	Statistic	Dyadic effects
<i>(a) Binary product attributes</i>		
Sum variable	$X_{ij} = x_i + x_j$	Attribute baseline effect
Matching variable	$X_{ij} = I \{x_i = x_j\}$	Homophily effect
<i>(b) Categorical product attributes</i>		
Matching variable	$X_{ij} = I \{x_i = x_j\}$	Homophily effect
<i>(c) Continuous product attributes (standardized)</i>		
Sum variable	$X_{ij} = x_i + x_j$	Attribute baseline effect
Difference variable	$X_{ij} = x_i - x_j $	Homophily effect
<i>(d) Non-product related attributes</i>		
Distance variable	$X_{ij} = \ x_i - x_j\ _2$	Homophily effect

(i) $I\{\cdot\}$ represents the indicator function; (ii) $|\cdot|$ represents the absolute-value norm on the 1-dimension space; (iii) $\|\cdot\|_2$ represents the L_2 -norm on the n -dimension Euclidian space.

to the same node have endogenous relations. That means the emergence of a link is often related to other links. The ERGM introduced by [49, 50] is well known for its capability in modeling the interdependence among links in social networks. For example, two people who have a common friend are more likely to be friends with each other too, and therefore the three-person friendship relations form a triangle structure. Specific *network configurations*, including edges, stars, triangles, and cycles, can be used to represent different types of interdependence. The ERGM interprets the global network structure as a collective self-organized emergence of various local network configurations. The logic underlying ERGM is that it considers an observed network, \mathbf{y} , as one specific realization from a set of possible random networks, \mathbf{Y} , following the distribution in the following equation [45]:

$$\Pr(\mathbf{Y} = \mathbf{y}) = \frac{\exp(\boldsymbol{\theta}^T \mathbf{g}(\mathbf{y}))}{\kappa(\boldsymbol{\theta})}, \quad (4)$$

where $\boldsymbol{\theta}$ is a vector of model parameters, $\mathbf{g}(\mathbf{y})$ is a vector of the network statistics and attributes, and $\kappa(\boldsymbol{\theta})$ is a normalizing quantity to ensure (4) is a proper probability distribution. Equation (4) suggests that the probability of observing any particular network is proportional to the exponent of a weighted combination of network characteristics: one statistic $g(\mathbf{y})$ is more likely to occur if the corresponding θ is positive. Note that, in ERGM, the network itself is a random variable and the probability is evaluated on the entire network instead of a link as in (3) for dyadic models. In brief, the advantages of using ERGM in the context of product coconsideration are threefold: (1) using *network configurations* to characterize the endogenous effects among coconsideration links, (2) providing various dyadic variables to model different types of exogenous impacts of the product attributes, and (3) integrating both exogenous attribute effects and endogenous network effects in a unified framework.

3.3.2. Exogenous Dyadic Variables and Endogenous Network Effects. The exogenous dyadic variables used both in the dyadic model and in ERGM allow the modeling of two types

of effects between a pair of nodes with specific variables: the baseline effects of the attributes and the homophily effects, that is, the similarity or difference between the attributes of two nodes [44, 51]. In the context of the product coconsideration network, the baseline effects examine whether products with a specific attribute are more likely to be coconsidered than products without that attribute; for example, imported car models could be more likely to be coconsidered as compared to domestic car models. The homophily effects examine whether two products with similar attributes tend to have a coconsideration link. For example, customers are more likely to consider and compare products with similar prices. The development of dyadic variables supports the study of inherent product competition beyond the understanding of customer preferences.

Table 1 summarizes the guidelines of creating dyadic variables for different types of attributes such as binary, categorical, and continuous. For the product attributes under (a)–(c), the strength of link X_{ij} is determined by the corresponding attributes x_i and x_j associated with the linked products. Beyond product attributes, we also introduce nonproduct related attributes (d). For example, customer demographics can be included in the model to allow the prediction of the impact of customers' associations/similarities on product coconsideration relations. To create a dyadic variables related to customers' attributes, multivariable association techniques, for example, joint correspondence analysis (JCA) [52], have been used to compute the similarity of the customer-related attributes as the distance between two product points (x_i and x_j) in a metric space. In this paper, we follow the method presented in [53] to develop two categories of distance variables, the distance of customer perceived characteristics and demographic distance. The customer perceived characteristics are user-proposed tags to indicate their perceptions of the products, such as youthful, sophisticated, and business-oriented. Customer demographics include income and family information of the user groups of each of the car models. The inclusion of customer associations through these distance-based dyadic variables is a unique feature of our network-modeling approach.

TABLE 2: Representative network characteristics of the generated coconsideration network.

Number of nodes	Number of links	Average degree	Average path length	Average local cluster coefficient
389	2,431	12.5	3.34	0.26

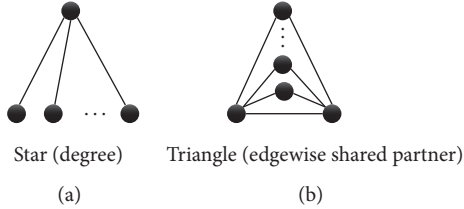


FIGURE 3: Two network configurations of coconsideration relations.

Different from the dyadic models that can only consider exogenous dyadic effects, the ERGM supports the modeling of product interdependence with *endogenous network effects*. In this paper, we are particularly interested in two *network configurations*, the star-type interdependence and triangle-type interdependence [1]. The star structures (Figure 3(a)) indicate that the probability of one focal product being coconsidered with others is conditional on the number of existing coconsideration relations of that focal product (e.g., the node on the top in the figure has three coconsideration links). A positive star effect suggests that a product is more likely to be coconsidered with another product if it is popular and already being coconsidered with many others. The triangle structures (Figure 3(b)) indicate that if two products are coconsidered with the same set of other products, they are more likely to be mutually coconsidered. Positive star effects could include stars with varying number of links (such as 2, 3, 4, 5, and perhaps many more). Likewise, a link could have many triangles by linking with varying number of nodes (1, 2, 3, 4, 5, and perhaps many more). Both star and triangular effects imply multiway product competition. To combine the effects of stars with multiple links and multiple triangles, we use two network configurations, the *geometrically weighted degrees* and the *geometrically weighted edgewise shared partner*, respectively [54].

4. Case Study: Modeling Vehicle Coconsideration Network

4.1. Application Context and Data Source. When considering and purchasing a vehicle, customers make decisions on car models (e.g., Ford Fusion versus Honda Accord), in part, based on their preferences for vehicle attributes (e.g., price, power, and make) and their demographics (e.g., income and age). To understand the effects of these factors on vehicles' coconsideration relations, we use data from a buyer survey in the 2013 China automarket. The dataset consists of about 50,000 new car buyers' responses to approximately 400 unique vehicle models. The survey covered a variety of questions, including respondent demographics, vehicle attributes, and customers' perceived vehicle characteristics. The respondents reported the car they purchased as well

as the primary and secondary alternatives they considered before making the final purchase. These responses are used to construct the vehicle coconsideration network. The vehicle attributes reported in the survey are verified by vehicle catalog databases.

4.2. Vehicle Coconsideration Network. Following the method discussed in Section 3.1, we construct a vehicle coconsideration network with $cutoff = 5$ which results in a network of 389 nodes and 2,431 binary links. A smaller *cutoff* generates a denser network but has similar analytical results. We have tested our models using *cutoff* at 1, 3, 5, and 7, respectively, and no significant changes in the trends of the model results are observed. Figure 4 shows an example of a partial vehicle coconsideration network with 11 car models. The node size is proportional to the degree, and colors indicate the clusters in which the vehicles are more likely to be coconsidered with each other. The number on each link is the *lift* value indicating the strength of the coconsideration.

Table 2 summarizes some descriptive network characteristics. For example, the average degree suggests that on average each vehicle has 12.5 coconsidered vehicles and indicates the overall intensity of competition in the market. The clustering coefficient (CC), on the other hand, measures the cohesion or segmentation of the vehicle market [44]. The average local CC at values of 0.26 indicates the strong cohesion embedded in the network, and vehicle models are frequently involved in multiway competition in the market. The descriptive network analysis facilitates the understanding of the automarket and provides guidelines on the selection of *network configurations* in ERGM.

4.3. Descriptive Statistics of the Independent Dyadic Variables. Many exogenous dyadic variables related to vehicle attributes, such as the difference and sum variables of car prices, engine power, fuel consumption, and matching variables of vehicle's market segments, and make origin, could change the patterns of coconsideration among the vehicle models. We use information gain analysis to select 12 most important dyadic variables among all 22 possible dyadic variables. The log transformation (base 2) is applied to the price and engine power variables to offset the effect of large outliers. Table 3 shows the descriptive statistics of the independent variables.

In total, six vehicle attributes, *import*, *price*, *engine power*, *fuel consumption*, *market segment*, and vehicles' *make origin*, are considered in the model. *Import* is a binary variable describing whether a car is imported ($import = 1$, 37.3%) or domestically produced ($import = 0$, 62.7%). As suggested in Table 1 and Section 3.3.2, we construct a sum dyadic variable of *import* to account for its baseline effect of whether each of the paired cars is both imported (value 2 for 13.90% of the pairs), one imported and one domestic (value 1 for 46.76%), or both domestic (value 0 for 39.34%). If the baseline

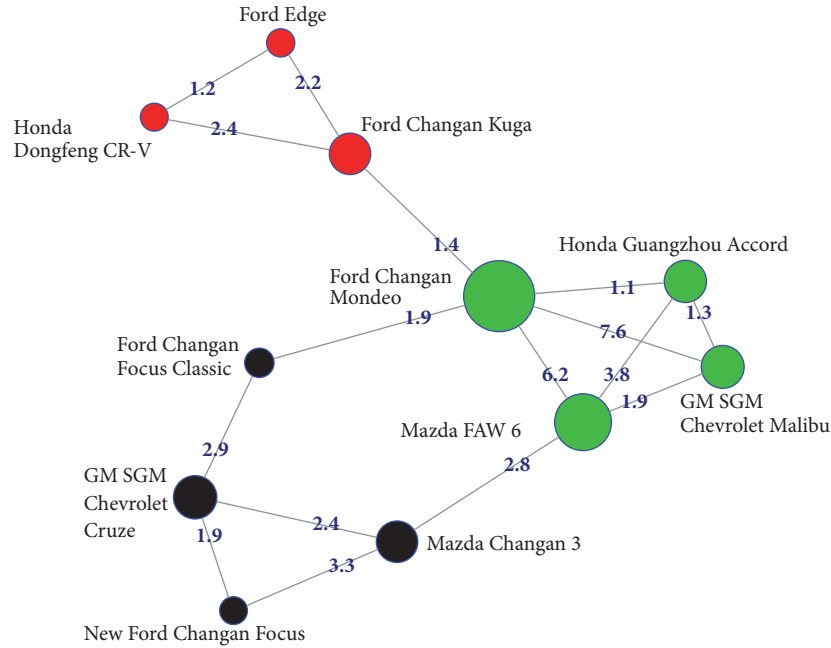


FIGURE 4: An example of partial vehicle coconsideration network.

TABLE 3: Descriptive statistics of independent variables for 389 car models in 2013.

	Mean (SD)	Min	Max
<i>Vehicle attributes</i>			
Import (binary)		145 import & 244 domestic	
Price (\log_2)	17.61 (1.34)	14.50	20.84
Power (\log_2)	7.27 (0.58)	5.25	8.76
Fuel consumption (per 100 BHP)	6.61 (1.62)	2.99	18.56
Market segment (categorical)		17 car segments	
Make origin (categorical)		13 American, 22 American-Chinese, 98 Chinese, 90 European, 50 European-Chinese, 31 Japanese, 54 Japanese-Chinese, 11 Korean, 20 Korean-Chinese	
<i>Vehicle attribute matching and difference</i>			
Market segment matching		10.1% pairs of cars coconsidered are in the same segment	
Make origin matching		16.5% pairs of cars coconsidered have the same make origin	
Price (\log_2) difference	1.53 (1.12)	0	6.34
Power (\log_2) difference	0.66 (0.49)	0	3.51
Fuel consumption difference	1.71 (1.52)	0	15.58
<i>Customer association</i>			
Distance of customers' perceived characteristics	0.20 (0.13)	0	1
Distance of customers' demographics	0.27 (0.16)	0	1

effect of the import attribute is positive, the coefficient of the sum variable of *import* should be positive as well, that is, the higher the sum value of the two car models, the more likely they are coconsidered together. Similarly, the sum variables of *price* (in RMB and transformed using \log_2) and *power* (in brake horsepower BHP and transformed using \log_2) describe the baseline effects of price and power on product coconsideration relations. We construct a variable,

fuel consumption, by dividing liters of gasoline each vehicle consumed per 100 kilometers over vehicle power (in 100 BHP). As such, the smaller this value is, the more fuel-efficient the car model is. The difference variables of price, power, and fuel consumption capture the homophily effects, which are used to test if the car models with similar attributes (smaller differences) are more likely to be coconsidered together.

TABLE 4: Estimated coefficients and odds ratios of the dyadic model and ERGM.

Input variables	Dyadic Model		ERGM	
	Est. coef.	Odds	Est. coef.	Odds
<i>Network configurations of product interdependence</i>				
Edge/Intercept	-14.36**	0.00	-13.71**	0.00
Star effect (inverse measure)			1.97**	7.20
Triangle effect			0.70**	2.01
<i>Baseline effects of vehicle attributes</i>				
Import	0.37**	1.45	0.11**	1.11
Price (log ₂)	-0.02	0.98	-0.007	1.01
Power (log ₂)	0.68**	1.97	0.35**	1.42
Fuel consumption (per 100 BHP)	0.19**	1.21	0.12**	1.13
<i>Homophily effects of vehicle attribute matching and difference</i>				
Market segment matching	1.38**	3.98	0.66**	1.94
Make origin matching	1.28**	3.60	0.53**	1.69
Price difference (log ₂)	-1.75**	0.17	-0.80**	0.45
Power difference (log ₂)	0.08	1.09	0.13	1.14
Fuel consumption difference	-0.08*	0.92	-0.07**	0.93
<i>Homophily effects of customer association</i>				
Distance of customers' perceived characteristics.	-0.42	0.66	-0.31	0.74
Distance of customers' demographics	-0.57**	0.56	-0.37*	0.69
<i>Model performance</i>				
Null deviance		104,618		
Bayesian Information Criterion (BIC)	16,005		14,021	

Note. * p -value < 0.01, ** p -value < 0.001.

The autoindustry is very competitive, so most car models have very clearly targeted customers and compete in a specific market segment. Since vehicle's *market segment* is a categorical variable, we use a dyadic matching variable in the model to investigate whether two cars from the same segment would affect their coconsideration patterns. The top 3 in all 17 segments in our sample are the C-Class Sedan (21.6% of car models), B-Class Sedan (11.3%), and Small Utility (11.1%). Similarly, *make origin* is also a categorical variable, and it describes the region where the car brand originates. Our dataset shows that 90, 31, 11, and 13 car models are made in Europe, Japan, South Korea, and the United States, respectively, 98 car models are produced in China with local brands and other local-foreign joint venture brands come from Europe (50), Japan (54), South Korea (20), and the United States (22). The matching variables of *market segments* and *make origins* are used to account for people's homophily behavior of comparing cars with the same brand and origin.

4.4. Model Implementation Using ERGM. Table 4 shows the estimated coefficients and corresponding odds ratios from fitting the dyadic and ERGM models. Other than the variables described above, the ERGM includes three additional variables associated with *network configurations*. The *edge* variable controls the number of links to ensure the estimated networks have the same density as the observed one. Conceptually, if we have no knowledge about the cars' attributes or their coconsideration relations, the *edge* estimates the likelihood that two cars will be coconsidered randomly, like

an intercept term in a regression or a "base rate". The *star effect* and *triangle effect* discussed in Section 3.3.2 are measured by *geometrically weighted degree* and the *geometrically weighted edgewise shared partner*, respectively. According to the results of the ERGM, most vehicle attributes, except the *price* baseline effect and *power* difference, are statistically significant (p value < 0.001) and therefore play important roles in vehicle coconsideration. For instance, two vehicles with smaller differences in price and fuel consumption are more likely to be coconsidered. If the price of one car model is twice the price of another car, their odds of coconsideration are only 45% of the odds of two cars with the same price. Similarly, one liter per 100 km per 100 BHP difference in fuel consumption leads to 93% of the odds of coconsideration compared to the cars with the same fuel consumption. For the matching of vehicle attributes, two vehicles in the same market segment are 1.94 times more likely to be coconsidered than the ones in different segments, and two vehicles with the same make origin are 1.69 times more likely to be coconsidered than the ones with different origins. Finally, the negative coefficient for the distance of customers' demographics shows that customers with different demographics are less likely to coconsider the same vehicle. In summary, the results show that customers are more likely to consider cars with similar perceived features, such as price, fuel consumption, market segment, and make origin.

As shown in Table 4, the coefficient of the *triangle effect* is 0.70 (p value < 0.001). The positive sign indicates that two vehicles coconsidered with the same set of vehicles

are more likely to be coconsidered with each other. It implies that a form of multiway grouping and comparison exists in customers' consideration decisions. That is, product alternatives in a person's consideration set are considered as the same time. On the other hand, the positive coefficient of the *star effect* (inversely measured by *geometrically weighted degree*) indicates that most of the cars tend to have a similar number of coconsideration links and there is an absence of a few cars that are much more likely to be coconsidered than others. With these endogenous network effects, the ERGM significantly improves the model fit compared to the dyadic model as indicated by the improvement of BIC from 16,005 to 14,021. In the next section, we perform a systematic comparative analysis to evaluate how well the simulated networks match the observed vehicle coconsideration network.

5. Model Comparison on Goodness of Fit

A goodness of fit (GOF) analysis is performed to compare the model fit of dyadic and ERGM models. Using the dyadic and ERGM models in (3) and (4), respectively, and based on the estimated parameters in Table 4, we compute the predicted probabilities of coconsideration between all pairs of vehicle models. The links with predicted probabilities higher than a threshold (e.g., 0.5) are considered as links that exist. Once the simulated networks are obtained from both models, we compare them against the observed 2013 coconsideration network at both the network level and the link level. The network-level evaluation uses the *spectral goodness of fit* (SGOF) metric [55], while the link level evaluation uses various accuracy measurements, such as *precision*, *recall*, and *F scores* (see Section 5.2 for more details).

5.1. Network-Level Comparison. Spectral goodness of fit (SGOF) is computed as follows:

$$\text{SGOF} = 1 - \frac{\text{E}\bar{\text{SD}}_{\text{obs,fitted}}}{\text{E}\bar{\text{SD}}_{\text{obs,null}}}, \quad (5)$$

where $\text{E}\bar{\text{SD}}_{\text{obs,fitted}}$ is the mean Euclidean spectral distance for the fitted model while $\text{E}\bar{\text{SD}}_{\text{obs,null}}$ is the mean Euclidean spectral distance for the null model, that is, the Erdős-Rényi (ER) random network in which each link has a fixed probability of being present or absent. Hence, SGOF measures the amount of the observed structures explained by a fitted model, expressed as a percent improvement over a null model. The Euclidean spectral distance computes the L_2 norm (also called Euclidean norm) of the error between the observed network and all k simulated networks, that is, $\|\epsilon_k\|$, where error ϵ is the absolute difference between the spectra of the observed network ($\hat{\lambda}^{\text{obs}}$) and that of the simulated network ($\hat{\lambda}^{\text{sim}}$), that is, $|\hat{\lambda}^{\text{obs}} - \hat{\lambda}^{\text{sim}}|$. Since the calculation of the spectra $\hat{\lambda}$ requires eigenvalues of the entire network's adjacent matrix, this evaluation is performed at the network level. When the fitted model exactly describes the data, SGOF reaches its maximum value 1. SGOF of zero means no improvement over the null model. The SGOF metric provides an overall comparison of different models. It is especially

TABLE 5: Spectral goodness of fit results of the dyadic model and ERGM.

	Dyadic model	ERGM
Mean SGOF (5th percentile, 95th percentile)	0.37 (0.31, 0.43)	0.63 (0.48, 0.76)

useful when a modeler is not clear about which network structural statistics are important in explaining the observed network. For example, in our coconsideration network, it is hard to tell which network metrics, such as the average path length or the average CC, are more important to the understanding of market structure. Under this circumstance, the SGOF provides a simple yet comprehensive evaluation. Table 5 lists the SGOF scores of both dyadic model and the ERGM. Based on 1,000 predicted networks from each model, the results of the mean, 5th, and 95th percentile of SGOF show that the ERGM significantly outperforms the dyadic model.

5.2. Link-Level Comparison. In addition to the network-level comparison, the predicted networks are also evaluated at the link level. We define a pair of vehicles with a coconsideration relation as *positive*, whereas the ones without links as *negative*. Therefore, the *true positive* (TP) is the number of links predicted as positive and also positive in the observed network; the *false positive* (FP) is the number of links predicted as positive but actually negative, that is, wrong predictions of positives. Similarly, the *true negative* (TN) is the number of links predicted as negative and observed as negative; the *false negative* (FN) is the number of links predicted as negative but observed as positive. Taking 0.5 as the threshold of predicted probability (as it is used in the logistic function), we calculate the following three metrics to evaluate the performance of prediction for both dyadic model and ERGM. *Precision* is the fraction of true positive predictions among all positive predictions; *recall* is the fraction of true positive predictions over all positive observations; *F score* is the harmonic mean of *precision* and *recall* (see Table 6 for the formulas). These metrics are adopted because each of them reflects the capability of the model from different perspectives. It could be the case where the model predicts many links (e.g., all links are predicted in extreme cases and FP is high) so that the *precision* is low and the *recall* is high, while another model could predict very few links that leads to high FN and therefore high *precision* and low *recall*. Therefore, using either *precision* or *recall* only practically reveals the model performance. Hence *F score* is often recommended as a fair measure because it considers both *precision* and *recall* and provides an average score. In this study, we use all three metrics together to provide a complete picture of the model performance.

As shown in Table 6, almost all performance metrics suggest that ERGM outperforms the dyadic model. In particular, the *recall* of ERGM is significantly higher than that of the dyadic model. The dyadic model is only able to predict about 4.2% of coconsideration, whereas the *recall* of the ERGM reach 31.1%. These results imply that the inclusion of product

TABLE 6: Results of various metrics for link-level comparison (predicted links based on threshold at 0.5).

Metrics	Dyadic model	ERGM
Precision = $\frac{TP}{(TP + FP)}$	0.594	0.543
Recall = $\frac{TP}{(TP + FN)}$	0.042	0.311
$F_\beta = \frac{(1 + \beta^2) \times \text{Precision} \times \text{Recall}}{(\beta^2 \times \text{Precision} + \text{Recall})}$	$F_{0.5} = 0.162$ $F_1 = 0.078$ $F_2 = 0.051$	$F_{0.5} = 0.473$ $F_1 = 0.396$ $F_2 = 0.340$

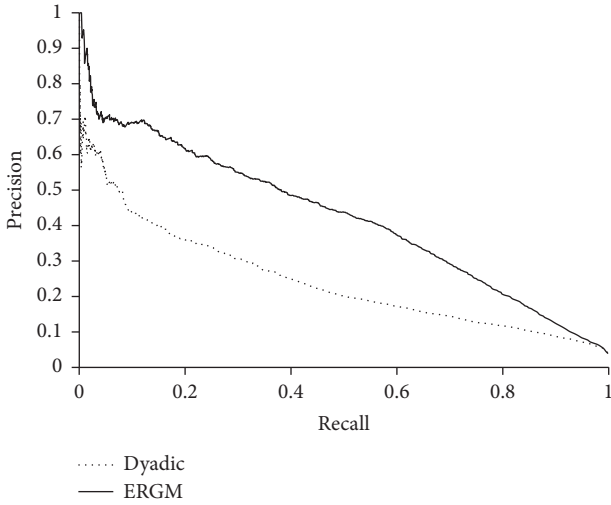


FIGURE 5: The precision-recall curve of the dyadic model and ERGM with random network benchmarked.

interdependence in ERGM indeed improves the model fit and better explains the observed product coconsideration relations. The only metric for which the dyadic model has a better value is the *precision*. At the threshold of probability equal to 0.5, the dyadic model only predict 170 links in total, and 101 of them are correct. The small denominator in the *precision* formula, that is, $TP + FP = 170$, produces a larger *precision*.

Since different thresholds of the predicted probability can affect the value of *precision* and *recall*, we evaluate the *precision-recall curve* [56] by altering the threshold from 0 to 1 to get a more comprehensive understanding. The model that has a larger area under the curve (AUC) performs better [57]. When evaluating binary classifiers in an imbalanced dataset (with many more cases of one value for a variable than the other), which is the case we face, Saito and Rehmsmeier [57] have demonstrated that the *precision-recall curve* is more informative than other threshold curves, such as the receiver operating characteristic (ROC) curve. Figure 5 shows that, for any given recall value, the *precision* of ERGM is strictly higher than that of the dyadic model and the ERGM outperforms the dyadic model in the full spectrum of the threshold of

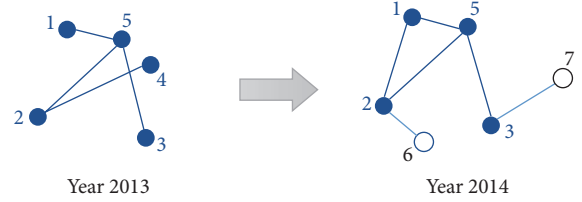


FIGURE 6: Illustration of the evolution of the coconsideration network.

probability (we studied the ROC curve and drew the same conclusions).

In summary, the comparisons at both the network level and the link level validate our hypothesis that the product interdependence, that is, the endogenous effect, plays a significant role in the formation of product coconsideration relations and hence the customers' consideration decisions. In the next section, we examine the predictive power of the two models.

6. Model Comparison on Predictability

In this section, we take a further step to compare the two models in terms of the predictability. We use the models developed with the 2013 dataset (i.e., the model coefficients shown in Table 4) to predict the vehicle coconsideration relations in the 2014 market. From an illustrative example in Figure 6, we can see that some car models (e.g., node 4) withdrew from the market in 2014, some new car models (e.g., node 6 and node 7) were introduced to the market, but most of the car models (e.g., nodes 1, 2, 3, and 5) remained in the 2014 market. In this paper, we focus on predicting the future coconsiderations among the overlapping car models in two consecutive years since the new models may introduce critical features not captured in the previous market, such as electric cars. In our study, 315 car models were available in both 2013 and 2014. Therefore, the task here is to predict whether each pair of cars among these 315 car models will be coconsidered in 2014 given their new vehicle attributes in 2014, the new customer demographics, existing market competition structures (The market competition structure is captured by the model coefficients of the three *network configurations* including the *edge*, *star effect*, and *triangle effect* discussed in Section 4.4.), and the model coefficients estimated based on the 2013 data.

Most pairs of cars have the same dyadic status (i.e., coconsidered or not) in 2013 and 2014. For example, if two car models were not coconsidered in 2013, customers continued to not coconsider these two in 2014. This case is not of interest because predicting nonexistence is much easier due to the imbalance nature of the network dataset and it does not provide new insights. Similarly, the persistent coconsideration in both 2013 and 2014 is also expected. Therefore, we focus on changes in two prediction scenarios: emergence and disappearance of coconsideration links from 2013 to 2014. As shown in Table 7, among 47,724 pairs of cars that were not coconsidered in 2013, 1,202 pairs were considered in 2014. The event of changing from not being

TABLE 7: Prediction scenarios of interest.

Prediction scenarios	Year 2013	Year 2014	Events of interest
Emergence of coconsideration	47,724 pairs of cars not coconsidered	1,202 pairs of new coconsideration	Yes
		46,522 no change	No
Disappearance of coconsideration	1,731 pairs of cars coconsidered	1,087 pairs no longer coconsidered	Yes
		644 no change	No

TABLE 8: The prediction precision and recall at the threshold of 0.5 in two prediction scenarios.

Prediction scenarios	Model	Number of events of interest ($TP + FN$)	Number of predictions ($TP + FP$)	Number of correct predictions (TP)	Prediction <i>precision</i>	Prediction <i>recall</i>	Prediction F_1
(1)	Dyadic	1202	36	9	0.250	0.0075	0.015
	ERGM		442	111	0.251	0.092	0.135
(2)	Dyadic	1087	1654	1076	0.651	0.990	0.785
	ERGM		1183	860	0.727	0.791	0.758

coconsidered to being coconsidered indicates the change of market competition potentially caused by the change of vehicle attributes such as prices. On the other hand, 1,731 pairs of cars were coconsidered in 2013 among the 315 car models, but 1,087 pairs were no longer coconsidered in 2014. We indicate the two cases in the last column of Table 7 where the predictions of 2014 network using 2013 model are the events of interest. The two “Yes” cases, predicting emerging coconsideration and disappearing coconsideration links, both represent the change of coconsideration status from 2013 to 2014 and are the positive outcomes of model predictions. Such predictions are more difficult (yet substantively more useful) to attain than the other two “No” cases of nochange. By testing both the dyadic and ERGM models, we examine which model had better predictive capability, assuming that the driving factors and customer preferences of coconsideration characterized by the model coefficients in Table 4 are unchanged from 2013 to 2014.

In both prediction scenarios, we input the new values of vehicle attributes and customer profile attributes from 2014 into the model. When using ERGM, characteristics of *network configurations* calculated based on the 2013 data also served as inputs for prediction. Once the models predict the probability of each pair of car models, we evaluate the performance metrics separately in two scenarios: (1) the *precision* and *recall* of predicting emerging coconsideration among the 47,724 pairs of not coconsidered car models, and (2) the *precision* and *recall* of predicting the disappearance of coconsideration among 1,731 pairs of cars coconsidered in 2013. The *precision* and *recall* of predictions are calculated similarly to the ones used in Section 5.2. The *precision* score is the ratio of the number of correctly predicted links (such as corrected prediction of emerging coconsideration or disappeared coconsideration) over the number of predictions a model makes. The *recall* score is the ratio of the number of correctly predicted links over the number of events of interest (true emerging coconsideration or disappeared coconsideration in 2014).

Table 8 shows the results of the prediction *precision* and *recall* calculated based on the predicted probability of 0.5 as the threshold in the two scenarios. To predict emerging coconsideration, the ERGM had much better performance than the dyadic model. Specifically, the dyadic model tends to be overtrained based on vehicle attributes and only predicts a small set of most likely links, that is, 9 of the 1,202 emerging new coconsideration relations. On the other hand, the ERGM predicted 111 (more than ten times) emerging coconsideration with the same *precision*. With the probability threshold of 0.5, the ERGM and dyadic model had similar differences in performance in predicting disappearing coconsideration links. Figure 7(b) shows that ERGM outperforms the dyadic model in almost all points of the *precision-recall* curve. In fact, the PR curves (Figure 7) show that ERGM at the entire range of the threshold outperforms the dyadic model in both prediction scenarios.

Therefore, we conclude that the ERGM has better predictability than the dyadic model. In addition to the GOF fitness test, the prediction test described above further validates our hypothesis that taking interdependencies in network modeling better explains the coconsideration network. In this particular case study, the analyses performed in both GOF and prediction analyses indicate that vehicles’ coconsideration relations are influenced by their existing competitions in the market.

7. Closing Comments

In this paper, we propose a network-based approach to study customer preferences in consideration decisions. Specifically, we apply the *lift* association metric to convert customers’ considerations into a product coconsideration network in which nodes present products and links represent coconsideration relations between products. With the created coconsideration networks, we adopt two network models, the dyadic model and the ERGM, to predict whether two products would have a coconsideration relation or not. Using vehicle design

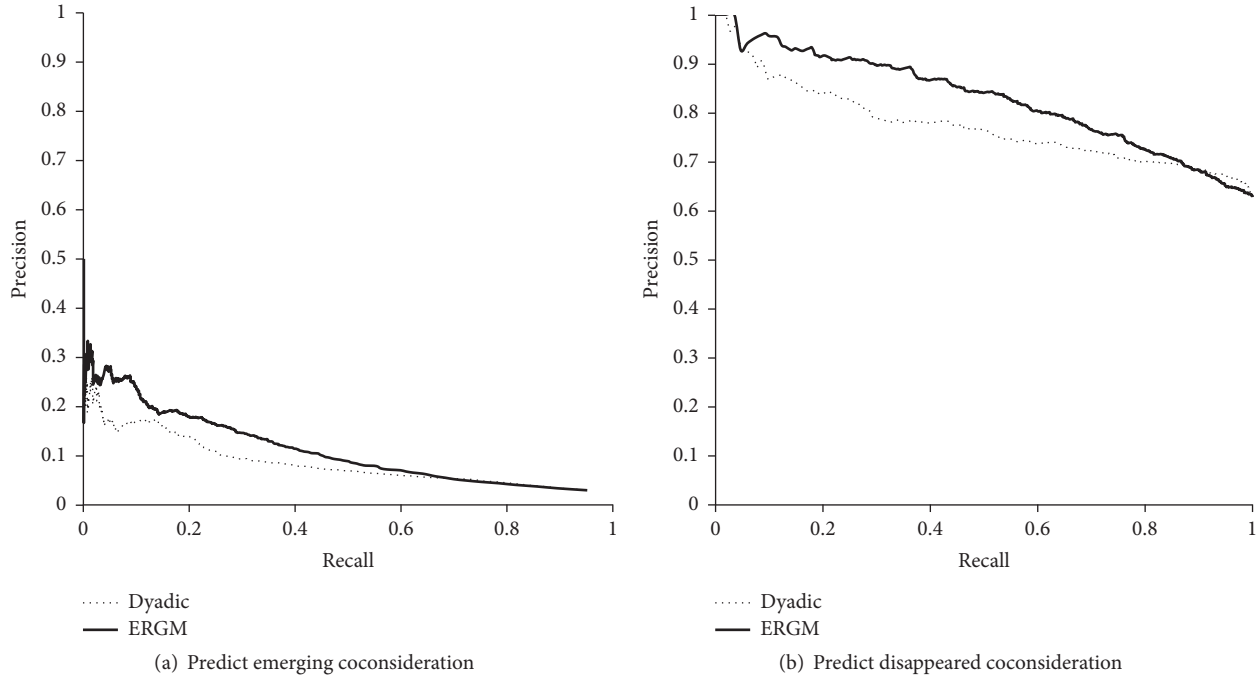


FIGURE 7: Prediction PR curves of dyadic model and ERGM in two prediction scenarios.

as a case study, we perform systematic studies to identify the significant factors influencing customers' coconsideration decisions. These factors include vehicle attributes (*price, power, fuel consumption, import, make origin, and market segment*), the similarity of customer demographics, and existing competition structures (i.e., the interdependence among coconsideration choices captured by *network configurations*). Statistical regressions are performed to obtain the estimated parameters of both models, and comparative analyses are performed to evaluate the models' goodness of fit and predictive power in the context of vehicle coconsideration networks. Our results show that the ERGM outperforms the dyadic model in both GOF tests and the prediction analyses. This paper makes two contributions relevant to engineering design: (a) a rigorous network-based analytical framework to study product coconsideration relations in support of engineering design decisions, and (b) a systematic evaluation framework for comparing different network-modeling techniques using GOF and prediction *precision* and *recall*.

This study provides three practical insights on coconsideration behavior in China automarket. First, the customers are price-driven when considering potential car models. Both models suggest significant homophily effects of vehicle prices and customer demographics in forming coconsideration links, that is, car models with similar prices and targeting to similar demographics such as income and family size are more likely to be considered in the same consideration set. However, the ERGM reveals much more influential drivers, such as the homophily effects of car segments and make origins. These findings confirm the internal clusters in the automarket. Second, the ERGM model suggests that there are

significantly fewer star structures but much more triangles in the coconsideration network. Beyond the impacts of the vehicle and customer attributes, ERGM also illustrates that car models that received an equal amount of consideration are likely to get involved in multiway coconsideration. Third, the model comparisons based on the GOF and prediction analyses demonstrate that an ERGM approach, which captures the interdependence of coconsideration, helps improve the prediction of product coconsiderations.

Finally, having an analytical model in this application context could boost future explorations including the what if scenario analysis that aims to forecast market responses under different settings of existing product attributes, as demonstrated in [44]. Since ERGM has a better model fit and predictability, it will help make more accurate projections on the future market trends and aid the prioritization of product features in satisfying customers' needs as well as support engineering design and product development. Future research should extend the network approach to a longitudinal weighted network-modeling framework, which not only predicts the existence of a link but also the strength of the coconsideration between car models in subsequent years. The weighted network models would help discover the nuance in different customers' consideration sets and therefore provide more insights into product design and market forecasting.

Disclosure

An early version of part of this work was presented in the 2017 International Conference on Engineering Design [58].

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper. Funding sources mentioned in Acknowledgments do not lead to any conflicts of interest regarding the publication of this manuscript.

Acknowledgments

The authors gratefully acknowledge the financial support from NSF CMMI-1436658 and Ford-Northwestern Alliance Project.

References

- [1] G. Robins, T. Snijders, P. Wang, M. Handcock, and P. Pattison, "Recent developments in exponential random graph (p^*) models for social networks," *Social Networks*, vol. 29, no. 2, pp. 192–215, 2007.
- [2] M. Wang, W. Chen, Y. Huang, N. S. Contractor, and Y. Fu, "A Multidimensional Network Approach for Modeling Customer-Product Relations in Engineering Design," in *Proceedings of the ASME 2015 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, American Society of Mechanical Engineers, Boston, Massachusetts, USA, 2015.
- [3] P. Wang, G. Robins, P. Pattison, and E. Lazega, "Exponential random graph models for multilevel networks," *Social Networks*, vol. 35, no. 1, pp. 96–115, 2013.
- [4] M. E. Sosa, S. D. Eppinger, and C. M. Rowles, "A network approach to define modularity of components in complex products," *Journal of Mechanical Design*, vol. 129, no. 11, pp. 1118–1129, 2007.
- [5] M. Sosa, J. Mihm, and T. Browning, "Degree distribution and quality in complex engineered systems," *Journal of Mechanical Design*, vol. 133, no. 10, Article ID 101008, 2011.
- [6] E. Byler, "Cultivating the growth of complex systems using emergent behaviours of engineering processes," in *Proceedings of the in International conference on complex systems: control and modeling*, Russian Academy of Sciences, 2000.
- [7] N. Contractor, P. R. Monge, and P. Leonardi, "Multidimensional networks and the dynamics of sociomateriality: Bringing technology inside the network," *International Journal of Communication*, vol. 5, pp. 682–720, 2011.
- [8] P. Cormier, E. Devendorf, and K. Lewis, "Optimal process architectures for distributed design using a social network model," in *Proceedings of the ASME 2012 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, IDETC/CIE 2012*, pp. 485–495, USA, August 2012.
- [9] W. Chen, C. Hoyle, and H. J. Wassenaar, "Decision-based design: Integrating consumer preferences into engineering design," *Decision-Based Design: Integrating Consumer Preferences into Engineering Design*, pp. 1–357, 2013.
- [10] J. J. Michalek, F. M. Feinberg, and P. Y. Papalambros, "Linking marketing and engineering product design decisions via analytical target cascading," *Journal of Product Innovation Management*, vol. 22, no. 1, pp. 42–62, 2005.
- [11] Z. Sha, K. Moolchandani, J. H. Panchal, and D. A. DeLaurentis, "Modeling Airlines' Decisions on City-Pair Route Selection Using Discrete Choice Models," *Journal of Air Transportation*, vol. 24, no. 3, pp. 63–73, 2016.
- [12] Z. Sha and J. H. Panchal, "Estimating local decision-making behavior in complex evolutionary systems," *Journal of Mechanical Design*, vol. 136, no. 6, Article ID 061003, 2014.
- [13] M. Wang and W. Chen, "A data-driven network analysis approach to predicting customer choice sets for choice modeling in engineering design," *Journal of Mechanical Design*, vol. 137, no. 7, Article ID 71409, 2015.
- [14] Z. Sha and J. H. Panchal, "Estimating linking preferences and behaviors of autonomous systems in the internet using a discrete choice model," in *Proceedings of the 2014 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2014*, pp. 1591–1597, usa, October 2014.
- [15] J. Hauser, M. Ding, and S. P. Gaskin, "Non-compensatory (and compensatory) models of consideration-set decisions," in *Proceedings of the in 2009 Sawtooth Software Conference Proceedings*, Sequim WA, 2009.
- [16] A. D. Shocker, M. Ben-Akiva, B. Boccara, and P. Nedungadi, "Consideration set influences on consumer decision-making and choice: Issues, models, and suggestions," *Marketing Letters*, vol. 2, no. 3, pp. 181–197, 1991.
- [17] J. R. Hauser and B. Wernerfelt, "An Evaluation Cost Model of Consideration Sets," *Journal of Consumer Research*, vol. 16, no. 4, p. 393, 1990.
- [18] P. Nedungadi, "Recall and Consumer Consideration Sets: Influencing Choice without Altering Brand Evaluations," *Journal of Consumer Research*, vol. 17, no. 3, p. 263, 1990.
- [19] R. K. Srivastava, M. I. Alpert, and A. D. Shocker, "A Customer-Oriented Approach for Determining Market Structures," *Journal of Marketing*, vol. 48, no. 2, p. 32, 1984.
- [20] S. Frederick, Automated choice heuristics.
- [21] W. Shao, *Consumer Decision-Making: An Empirical Exploration of Multi-Phased Decision Processes*, Griffith University Australia, 2006.
- [22] M. Yee, E. Dahan, J. R. Hauser, and J. Orlin, "Greedoid-based noncompensatory inference," *Marketing Science*, vol. 26, no. 4, pp. 532–549, 2007.
- [23] J. R. Hauser, O. Toubia, T. Evgeniou, R. Befurt, and D. Dzyabura, "Disjunctions of conjunctions, cognitive simplicity, and consideration sets," *Journal of Marketing Research*, vol. 47, no. 3, pp. 485–496, 2010.
- [24] T. J. Gilbride and G. M. Allenby, "Estimating heterogeneous EBA and economic screening rule choice models," *Marketing Science*, vol. 25, no. 5, pp. 494–509, 2006.
- [25] R. L. Andrews and T. C. Srinivasan, "Studying consideration effects in empirical choice models using scanner panel data," *Journal of Marketing Research*, vol. 32, no. 1, p. 30, 1995.
- [26] J. Chiang, S. Chib, and C. Narasimhan, "Markov chain Monte Carlo and models of consideration set and parameter heterogeneity," *Journal of Econometrics*, vol. 89, no. 1-2, pp. 223–248, 1998.
- [27] T. Erdem and J. Swait, "Brand credibility, brand consideration, and choice," *Journal of Consumer Research*, vol. 31, no. 1, pp. 191–198, 2004.
- [28] J. Swait, "A non-compensatory choice model incorporating attribute cutoffs," *Transportation Research Part B: Methodological*, vol. 35, no. 10, pp. 903–928, 2001.
- [29] T. J. Gilbride and G. M. Allenby, "A choice model with conjunctive, disjunctive, and compensatory screening rules," *Marketing Science*, vol. 23, no. 3, pp. 391–406, 2004.

- [30] E. Boros, P. L. Hammer, T. Ibaraki, A. Kogan, E. Mayoraz, and I. Muchnik, "An implementation of logical analysis of data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 12, no. 2, pp. 292–306, 2000.
- [31] M. Ding, "An incentive-aligned mechanism for conjoint analysis," *Journal of Marketing Research*, vol. 44, no. 2, pp. 214–223, 2007.
- [32] Z. Sha, V. Saeger, M. Wang, Y. Fu, and W. Chen, "Analyzing Customer Preference to Product Optional Features in Supporting Product Configuration," *SAE International Journal of Materials and Manufacturing*, vol. 10, no. 3, 2017.
- [33] K. Eliaz and R. Spiegler, "Consideration sets and competitive marketing," *Review of Economic Studies*, vol. 78, no. 1, pp. 235–262, 2011.
- [34] P. Resnick and H. R. Varian, "Recommender systems," *Communications of the ACM*, vol. 40, no. 3, pp. 56–58, 1997.
- [35] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734–749, 2005.
- [36] T. Zhoua, Z. Kuscsik, J. Liu, M. Medo, J. R. Wakeling, and Y. Zhang, "Solving the apparent diversity-accuracy dilemma of recommender systems," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 10, pp. 4511–4515, 2010.
- [37] F. Yu, A. Zeng, S. Gillard, and M. Medo, "Network-based recommendation algorithms: a review," *Physica A: Statistical Mechanics and its Applications*, vol. 452, pp. 192–208, 2016.
- [38] L. Lü, M. Medo, C. H. Yeung, Y. Zhang, Z. Zhang, and T. Zhou, "Recommender systems," *Physics Reports*, vol. 519, no. 1, pp. 1–49, 2012.
- [39] T. Zhou, J. Ren, M. Medo, and Y. Zhang, "Bipartite network projection and personal recommendation," *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 76, no. 4, Article ID 046115, 2007.
- [40] M. J. Pazzani and D. Billsus, "Content-based recommendation systems," in *The Adaptive Web: Methods and Strategies of Web Personalization*, P. Brusilovsky, A. Kobsa, and W. Nejdl, Eds., pp. 325–341, Springer, Berlin, Germany, 2007.
- [41] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. 4–5, pp. 993–1022, 2003.
- [42] A. Fiasconaro, M. Tumminello, V. Nicosia, V. Latora, and R. N. Mantegna, "Hybrid recommendation methods in complex networks," *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 92, no. 1, Article ID 012811, 2015.
- [43] J. S. Fu et al., "Modeling Customer Choice Preferences in Engineering Design Using Bipartite Network Analysis," in *Proceedings of the in 2017 ASME International Design Engineering Technical Conferences & Computers and Information in Engineering Conference*, ASME, Cleveland, OH, USA, 2017.
- [44] M. Wang, Z. Sha, Y. Huang, N. Contractor, Y. Fu, and W. Chen, "Forecasting technological impacts on customers' co-consideration behaviors: A data-driven network analysis approach," in *Proceedings of the ASME 2016 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, IDETC/CIE 2016*, USA, August 2016.
- [45] G. Robins, P. Pattison, Y. Kalish, and D. Lusher, "An introduction to exponential random graph (p*) models for social networks," *Social Networks*, vol. 29, no. 2, pp. 173–191, 2007.
- [46] M. Shumate and E. T. Palazzolo, "Exponential random graph (p*) models as a method for social network analysis in communication research," *Communication Methods and Measures*, vol. 4, no. 4, pp. 341–371, 2010.
- [47] S. Tufféry, "Data Mining and Statistics for Decision Making," *Data Mining and Statistics for Decision Making*, 2011.
- [48] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Computational linguistics*, vol. 16, no. 1, pp. 22–29, 1990.
- [49] O. Frank and D. Strauss, "Markov graphs," *Journal of the American Statistical Association*, vol. 81, no. 395, pp. 832–842, 1986.
- [50] S. Wasserman and P. Pattison, "Logit models and logistic regressions for social networks: I. An introduction to markov graphs and p," *Psychometrika*, vol. 61, no. 3, pp. 401–425, 1996.
- [51] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: homophily in social networks," *Annual Review of Sociology*, vol. 27, pp. 415–444, 2001.
- [52] M. Greenacre, *Correspondence analysis in practice*, CRC press, 2017.
- [53] M. Wang et al., "A Network Approach for Understanding and Analyzing Product Co-Consideration Relations in Engineering Design," in *Proceedings of the DESIGN 2016 14th International Design Conference*, 2016.
- [54] D. R. Hunter, "Curved exponential family models for social networks," *Social Networks*, vol. 29, no. 2, pp. 216–230, 2007.
- [55] J. Shore and B. Lubin, "Spectral goodness of fit for network models," *Social Networks*, vol. 43, pp. 16–27, 2015.
- [56] D. M. Powers, *Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation*, 2011.
- [57] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, Article ID e0118432, 2015.
- [58] Z. Sha et al., "Modeling Product Co-Consideration Relations: A Comparative Study of Two Network Models," in *Proceedings of the 21st International Conference on Engineering Design, ICED17*, 2017.

Research Article

Bipartisanship Breakdown, Functional Networks, and Forensic Analysis in Spanish 2015 and 2016 National Elections

Juan Fernández-Gracia^{1,2} and Lucas Lacasa ³

¹Harvard T.H. Chan School of Public Health, Harvard University, 677 Huntington Ave, Boston, MA 02115, USA

²Instituto de Física Interdisciplinar y Sistemas Complejos (IFISC), CSIC-UIB, Mallorca, Spain

³School of Mathematical Sciences, Queen Mary University of London, Mile End Road, London E14NS, UK

Correspondence should be addressed to Lucas Lacasa; l.lacasa@qmul.ac.uk

Received 24 October 2017; Accepted 4 December 2017; Published 24 January 2018

Academic Editor: Gerard Olivar-Tost

Copyright © 2018 Juan Fernández-Gracia and Lucas Lacasa. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

We present a social network and forensic analysis of the vote counts of Spanish national elections that took place in December 2015 and their sequel in June 2016. We initially consider the phenomenon of bipartisanship breakdown by analyzing spatial distributions of several bipartisanship indices. We find that such breakdown is more prominently close to cosmopolite and largely populated areas and less important in rural areas where bipartisanship still prevails, and its evolution mildly consolidates in the 2016 round, with some evidence of bipartisanship reinforcement which we hypothesize to be due to psychological mechanisms of risk aversion. Subsequently, a functional network analysis detects an effective partition of municipalities which remarkably coincides with the first-level political and administrative division of autonomous communities. Finally, we explore to which extent vote data are faithful by applying forensic techniques to vote statistics. Results based on deviation from Benford's law are mixed and vary across different levels of aggregation. As a complementary metric, we further explore the cooccurring statistics of vote share and turnout, finding a mild tendency in the clusters of the conservative party to smear out towards the area of high turnout and vote share, what has been previously interpreted as a possible sign of incremental fraud.

1. Introduction and Datasets

In the last decade and in parallel with the improvement of computational resources and the possibility of accessing, storing and manipulating massive digital records easily, the political science community has engaged with the task of producing quantitative and systematic methods to detect irregularities in electoral results [1]. In this work, we analyze the vote count statistics obtained in the Spanish national elections that took place in December 2015 as well as in their sequel of June 2016. Since the end of 2014, the emergence of new parties such as the antiausterity Podemos and the rise of other ones such as Ciudadanos (Cs) challenged an already decadent bipartisanship system, as was evidenced by the highly fragmented total vote share in 2015. These results further defined a new type of political equilibrium in Spain, where the quest for alliances across parties was required to form a workable majority. Unfortunately, this situation was

not achieved and the parliament was unable to build the necessary coalitions to make such a workable majority, what triggered the onset of new elections only six months after the previous ones, in June 2016. These special and unique conditions, together with the fact that the polls and electoral surveys preceding and on the day of the elections, showed an unusually high discrepancy with the actual results motivating the use of some of the recently developed techniques for elections forensic analysis to scrutinize any source of irregularity in these elections.

High resolution vote count data (at several levels of aggregation down to the level of municipalities) have been extracted from the official webpage of the Ministerio Del Interior [2] (Spanish Ministry of Home Affairs) for both 2015 and 2016 elections. For concreteness we have focused on vote counts on congress and discarded senate (see Figure 1 for a guide of the type of data available from the ministry of home affairs website).

Arteixo

Datos de las 00:15 del 27 de junio de 2016
Escrutado 100%

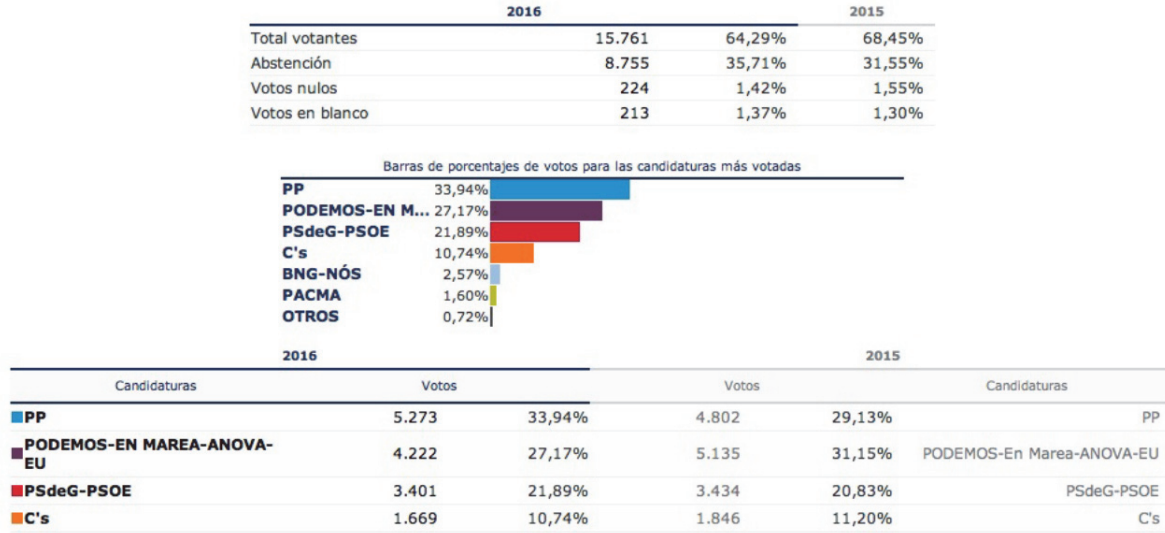


FIGURE 1: Sample municipality (Arteixo) along with vote count statistics, as reported in the Ministerio del Interior official webpage [2].

The high quality of the dataset under study allows us to address important social and political questions in a scientific way. In this work, we have considered three specific questions, namely, (i) can we confirm that the bipartisanship system is challenged? (ii) Can a social network analysis reveal quantitative information on the voting profiles and similarities across municipalities and regions? And (iii) can we scrutinize these data against forensic techniques? And if so, is there any evidence of fraud?

To address the first question, we will define a set of bipartisanship indices and will explore the spatial distribution of these over the Iberian peninsula of Spain (at the fine-grained level of municipalities), while, for the second question, we will build on state-of-the-art community detection methods (Infomap) applied on a functional-like network of municipalities extracted via cosine similarity.

The third question will be addressed via two different studies. The first one addresses the deviation or conformance of vote counts statistics to the so-called Benford's law [3, 4] that predicts that the first significant digits in some datasets (including vote counts) should follow an inverse-logarithmic distribution. The rationale for this analysis is that statistically significant deviations between the empirical distribution and the theoretical one point us towards electoral irregularities. These irregularities might in turn be due either to unintentional mismanagement of the voting process and/or to fraud. This type of analysis only flags the existence of such irregularities and gives no judgment on what was the cause for such irregularity. To complement this study, we then explore the presence and detection of sources of incremental and extreme fraud from the cooccurring statistics of vote and turnout numbers, following a recent study [5].

The rest of the paper goes as follows: in Section 2, we introduce similarity indices and explore the spatial distribution of these. Having access to the 2015 and 2016 election statistics, we will be able to explore the social effect of distrust and the possible longitudinal progression of bipartisanship breakdown. In Section 3, we perform the social network analysis of the data, creating functional networks via cosine similarity computed on the vote profile at the level of municipalities. Then, in Section 4, we focus on forensic methods. We introduce Benford's law along with the precise types of statistical tests that have been proposed in the realm of election forensics, and we present the results obtained from these tests for both the December 2015 and June 2016 elections at three different levels of aggregation. The main results and interpretations on this first study are reported in this section and additional material and analysis are shown in Appendix. In this section, we also present the second forensic analysis that addresses the cooccurring statistics of vote and turnout numbers. Finally, in Section 5, we provide some discussion and conclude.

2. Bipartisanship Indices

Vote counts can be aggregated at several spatial levels (municipalities, precincts, etc.). In general, for a specific region (e.g., a given municipality), vote data consist of a vector of vote percentage $\mathbf{v} = (v_1, v_2, \dots)$, where v_i is the percentage of votes to party i , and $\sum_{i=1}^N v_i = 1$, N being the total number of parties with representation in that region. Whereas in a majority of municipalities the main national-wide parties PP, PSOE, Cs, and Podemos have representation, other smaller, regional (or otherwise) parties also appear with different frequency in the

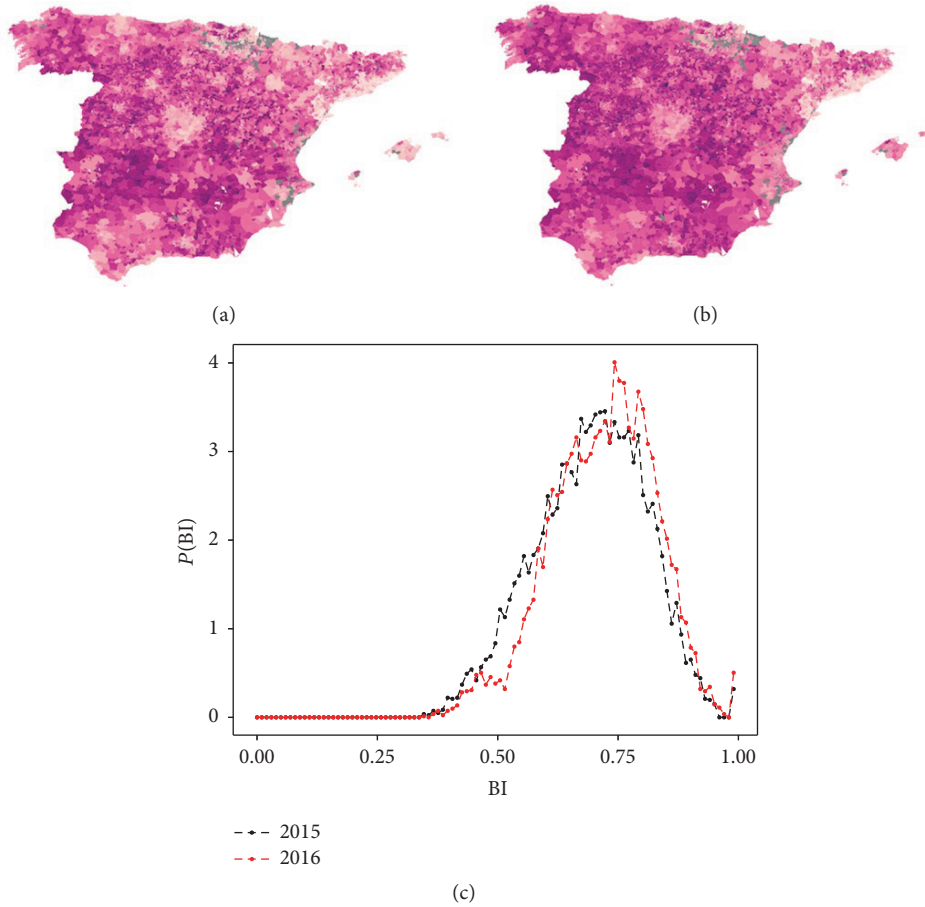


FIGURE 2: (a, b) Heat map of the bipartisanship index (BI, see the text) at the municipality level (the darker and the larger the BI is, the more bipartisan the system is) for 2015 elections (a) and 2016 elections (b). One can see that overall bipartisanship decreases from 2015 to 2016 and that bipartisanship breakdown is more acute closer to important, cosmopolite cities (e.g., Madrid and Barcelona). (c) Frequency histogram of BI for 2015 and 2016. In 2015, the distribution has a clear Gaussian shape and such shape is slightly perturbed in 2016.

different municipalities. In other words, N (total number of parties) might fluctuate from municipality to municipality.

Bipartisanship Index (BI). As a crude metric, we initially define a *bipartisanship index* (BI_i) of a given municipality (resp., precinct, etc.) i as the sum of the vote percentage ratio (between 0 and 1) of the two most-voted parties in i . For a pure bipartisanship region i ($N = 2$), BI_i approaches 1, whereas in the ideal case of a multiparty region, BI would approach its minimum value $2/N$. This metric allows us to compare the relative level of bipartisanship across regions.

In Figure 2, we provide a spatial heat map of Spain where we plot BI_i for each municipality $i = 1, 2, \dots, 8215$, for the 2015 and 2016 elections data, respectively (note that, for illustration constraints, Canary Islands are not represented here, but we shall emphasize that data and results there are qualitatively equivalent to those obtained for the Iberian peninsula and Balearic Islands). First, we can observe that there is a clear tendency towards relatively lower bipartisanship in areas that correspond to highly populated regions, for example, Madrid and Catalonia. This finding is well aligned with the social observation that the process of bipartisanship

breakdown, as other social changes initially develop close to important cosmopolite cities and then percolate to more rural areas. A second interesting finding is that from 2015 to 2016 there is a *stalling* in the bipartisanship breakdown, and its average index over all Spanish municipalities even slightly increases from $\langle BI \rangle = 0.70 \pm 0.12$ in 2015 to $\langle BI \rangle = 0.73 \pm 0.11$ in 2016 (this small increase is however within error bars so one cannot rule out this being a statistical artifact). A possible sociological interpretation is the following: after the 2015 elections, parliament was unable to build the necessary coalitions to make a workable majority, and this was the trigger to the new elections only six months after the previous ones, in June 2016. As society is averse to the uncertainty generated by frustrated elections, risk aversion might have stalled the overall inertia that a priori was driving the bipartisanship breakdown, and the fear of not being able to find workable majorities might have forced some voters in the most conservative regions to turn back to a bipartisanship strategy that indeed guarantees those much-needed majorities.

At a regional level, we can observe that this phenomenon is highly heterogeneous: whereas there is a very acute trend

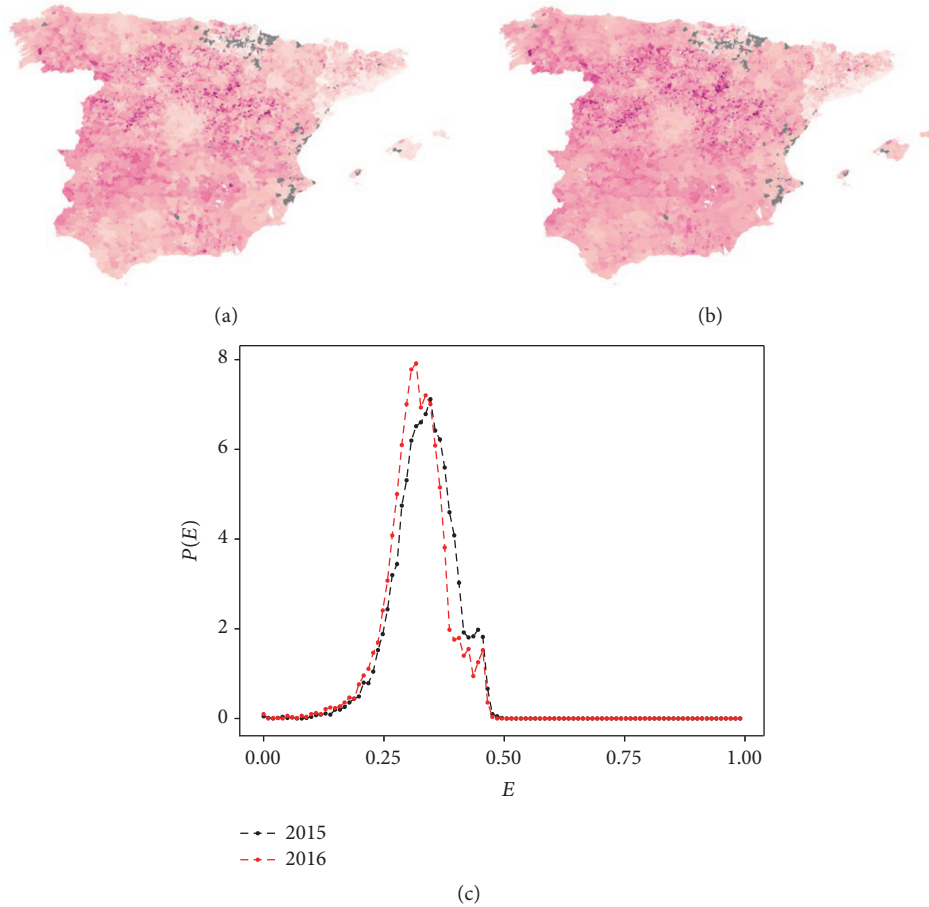


FIGURE 3: (a, b) Heat maps of the entropy index (E , see the text) at the municipality level (the darker and the lower the entropy is, the more the bipartisan system is) for 2015 elections (a) and 2016 elections (b). Entropy is larger (leading to more heterogeneous vote and less bipartisanship) closer to cosmopolite cities (e.g., Madrid and Barcelona). Overall, this index is approximately constant over time ($\langle E \rangle = 0.34 \pm 0.06$ in 2015 and $\langle E \rangle = 0.33 \pm 0.06$ in 2016), similar to the results obtained using the bipartisanship index.

towards bipartisanship breakdown consolidation in specific regions such as the autonomous community of Valencia, in other regions such as the autonomous community of Galicia the trend is pretty much the opposite.

In Figure 2(c) we plot the frequency histogram of bipartisanship indices, for both 2015 and 2016. For 2015, a clear Gaussian-like shape emerges, and such shape is slightly perturbed for the 2016 case. In the latter case, we observe weird peaks and pits emerging in the distribution leaving a trace of wild fluctuations which are not present in the 2015 statistics. Comparing both histograms we can perceive a subtle shift towards higher values of BI.

Entropy Index (E). Intuitively, bipartisanship tends to accumulate vote percentage in a few parties, whereas multiparty tends to spread out votes across parties. Following this argument, in order to quantify more precisely the concentration of vote percentages over the whole set of parties, one can define an *entropy index* for region i as

$$E_i = -\frac{\sum_{j=1}^N v_j \log v_j}{\log N}. \quad (1)$$

This quantity is bounded in $0 \leq E_i \leq 1$, reaching the minimum for uniparty (a single party gets 100% of the votes in the region) and reaching its maximum for multiparty (all N parties get the same percentage of votes).

Qualitatively, we have found similar results when exploring spatial distributions of the entropy index E_i to those obtained with the more crude quantifier BI (see Figure 3).

Diversity Index. Inspired by ecological metrics [6], here we further define the diversity index N_{eff} of a given region as the effective number of political parties. This refers to the number of “equally voted parties” needed to obtain the same mean proportional parties vote percentage as that observed in the dataset (where all parties may not be equally abundant):

$$N_{\text{eff}} = \exp(E \log N), \quad (2)$$

where E is the entropy index as defined above. In other words, N_{eff} counts, assuming that effective parties all get the same number of votes, the number of such effective parties one would need to find the same entropy index as found by computing E to the vote statistics. In Figures 4(a) and 4(b)

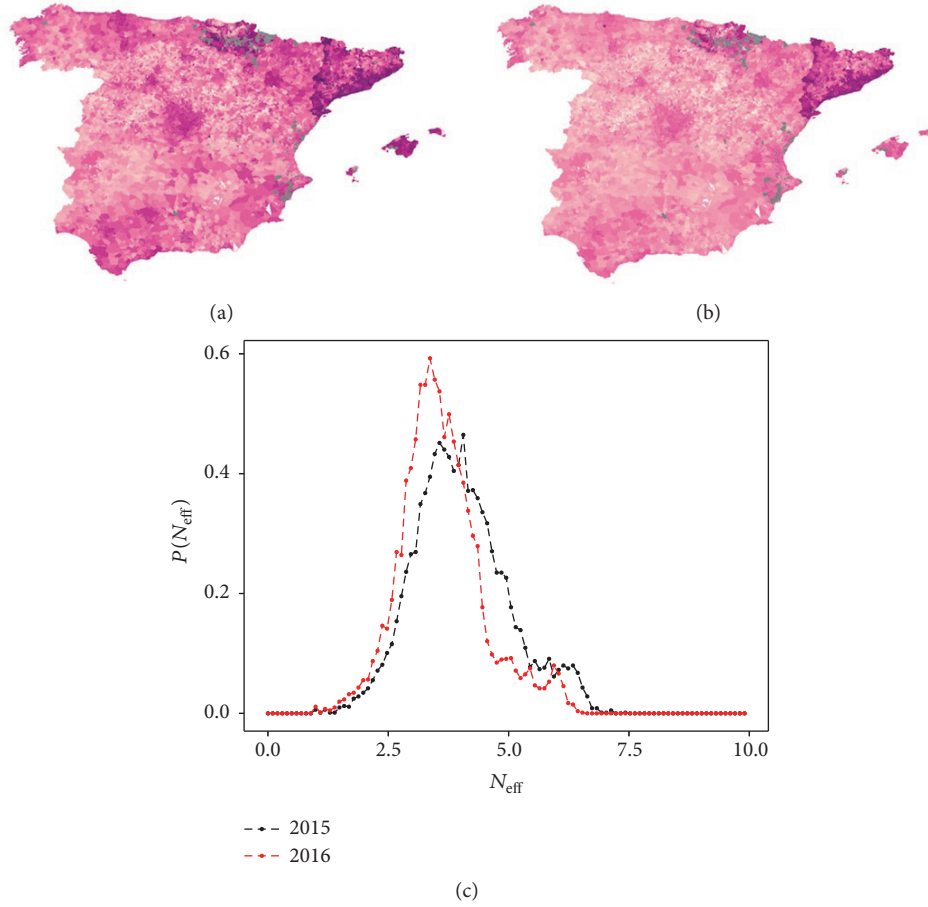


FIGURE 4: (a, b) Heat map of the diversity index (N_{eff} , see the text) at the municipality level (the darker and the larger N_{eff} is, the less bipartisan the system is) for 2015 elections (a) and 2016 elections (b). One can see that overall the system is more diverse in 2015 and its diversity decreases in 2016. This result is in consonance with bipartisanship breakdown stalling, probably due to risk aversion that generates the uncertainty associated with not obtaining workable majorities in parliament. (c) Frequency histogram of N_{eff} for 2015 and 2016, highlighting an initially diverse voting ecosystem with $\langle N_{\text{eff}} \rangle = 4.1$ (over four relevant parties on average) in 2015, and a clear shift towards a smaller diversity in 2016.

we plot such index for 2015 (a) and 2016 (b) elections. We clearly observe two important stylized facts: namely, (i) as already found in BI and E , there is a clear separation between regions close to cosmopolite and largely populated cities, whose diversity index tends to be large, entailing a high number of effective parties at play, and regions which are typically less populated (rural areas) with lower diversity index. (ii) At odds with BI, the overall index evidences a marked decrease between 2015 and 2016, visually observed by a drift in heat map towards lighter color (i.e., lower values of diversity) and a change in the diversity distribution (Figure 4(c)).

3. Functional Network Analysis

In this section, we make use of tools from Network Theory [7] to explore the vote similarity across regions. Attached to each municipality, we consider the vote percentage vector \mathbf{v} defined above. We measure the similarity between the voting

statistics of two municipalities i and j via the so-called cosine similarity:

$$S_{ij} = \frac{\langle \mathbf{v}^{(i)}, \mathbf{v}^{(j)} \rangle}{\|\mathbf{v}^{(i)}\| \cdot \|\mathbf{v}^{(j)}\|}, \quad (3)$$

where $\langle \cdot, \cdot \rangle$ is the standard scalar product and $\|\cdot\|$ is the ℓ_2 norm. By construction, $0 \leq S_{ij} \leq 1$, where complete similarity ($S_{ij} = 1$) is reached when the vote statistics are identical and null similarity is reached when $\mathbf{v}^{(i)} \perp \mathbf{v}^{(j)}$, that is, when a nonnull percentage to a given party in one municipality is always matched to a null percentage in the other municipality.

The matrix $\mathbf{S} = \{S_{ij}\}$ naturally defines a fully connected weighted network where nodes are municipalities, and every pair of nodes i and j are linked with an edge with weight S_{ij} . In our database, we have a total number of about $8 \cdot 10^3$ municipalities, hence a fully connected network of about $8 \cdot 10^3$ nodes and $64 \cdot 10^6$ edges. In Figure 5 we plot the edge's weight probability density $P(S)$. As expected, weights are concentrated in a region relatively close to the maximum

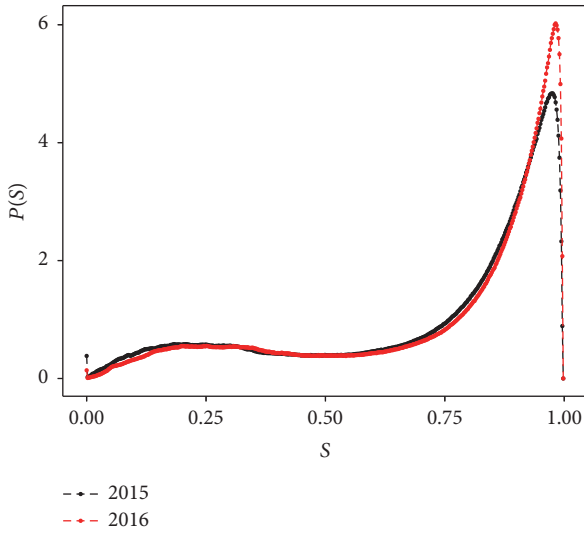


FIGURE 5: Estimated probability density of similarity values $P(S)$ for the fully connected, weighted, similarity matrix S .

value, something that can be justified already by noting that, in most of the municipalities, all four parties PP, PSOE, Cs, and Podemos have a nonnull vote percentage. To get some insight beyond this simple statistic, we now run an algorithm to detect communities, that is, large groups of nodes which are similar to each other and less similar to the nodes in the other groups. To do this, we run Infomap [8, 9] on the undirected and weighted network after a simple thresholding is performed on S : for a given node i , we only conserve the largest similarity weights S_{ij} such that all nodes at least have degree $k = 10$. In other words, we perform a parallel pruning on the edges, starting from those with smaller weights, and we prune the network up to the point where we cannot prune anymore if we want to make sure every node has a degree $k \geq 10$. The first level of Infomap reveals a total of 14 communities in 2015 and 16 communities in 2016. In Figure 6 we plot a spatial projection of the network, where nodes are municipalities. For exposition reasons, edges are not plotted in the figure as otherwise the image would not carry much information. Nodes belonging to the same network community are colored equally. Resemblance between network communities obtained via Infomap in our functional network and actual Spanish autonomous communities is remarkable, and such similitude is more acute in 2015, where bipartisanship breakdown is slightly more pronounced, than in 2016, where bipartisanship slightly reduced, as reported by the diversity index, and this has the effect of fuzzing up the relation between network communities and autonomous community divisions. Interestingly, one finds what we could call stable and compact autonomous communities (those which have a well-defined, stable over time and cohesive functional community counterpart): Catalonia, Madrid, Basque Country, and Navarra, whereas other set of communities have a clear counterpart in 2015 but a fuzzier one in 2016 (e.g., Comunitat Valencia, Andalucia, and Murcia). Another set of autonomous communities present a highly heterogeneous voting profiles and do not present any clear

functional community counterpart. Theoretical digressions and insights that could give a sociopolitical justification for this classification are left as an open question for future work.

4. Forensic Analysis

4.1. Benford's Law. The first significant digit (or leading digit) of a number is defined as its nonzero leftmost digit (e.g., the leading digit of 123 is 1 whereas the leading digit of 0.025 is 2). The so-called Benford's law is an empirical statistical law stating that in particular types of numeric datasets the probability of finding an entry whose first significant digit is d decays logarithmically as

$$P(d) = \log_{10} \left(1 + \frac{1}{d} \right), \quad (4)$$

where \log_{10} stands here for the decimal logarithm (note that trivially $\sum_{d=1}^9 P(d) = 1$). Perhaps counterintuitively, this law is quite different from the expected distribution arising from an uncorrelated random process (e.g., coin tossing or extracting numbers at random from an urn) which would yield a uniform distribution where every leading digit would be equally likely to appear. The logarithmically decaying shape given in (4) was empirically found first in 1881 by astronomer Simon Newcomb and later popularized and exhaustively studied by Benford [10]. Empirical datasets that comply to Benford's law emerge in disparate places as for stock prizes or physical constants, and some mathematical sequences such as binomial arrays or some geometric sequences have been shown to conform to Benford. A possible origin of this law has been rigorously explained by Hill [11], who proved a central limit-type theorem by which random entries picked from random distributions form a sequence whose leading digit distribution converges to Benford's law. Another explanation comes from the theory of multiplicative processes, as it is well known that power-law distributed stochastic processes follow Benford's law for the specific case of a density $1/x$ (see [12] and references therein for details). In practice, this law is expected to emerge in a range of empirical datasets where part or all of the following criteria hold: (i) the data ranges a broad interval encompassing several orders of magnitude rather uniformly, (ii) the data are the outcome of different random processes with different probability densities, and (iii) the data are the result of one or several multiplicative processes.

Mainly advocated by Nigrini [3], the application of Benford's law to detect fraud and irregularities, by observing anomalous and statistically significant deviations from (4) for datasets which otherwise should conform to that distribution, has become popular in recent years, and from now on we quote this a IBL test. Mansilla [13] and Roukema [14] applied this methodology to assess Mexican and Iranian vote count results, respectively. On the other hand, Mebane [4] advocates instead to look at the second significant digit (which follows an extended version of Benford's law [15]) and argues that the frequencies of election vote counts at precinct level approximate a Benford distribution for the second digit, and

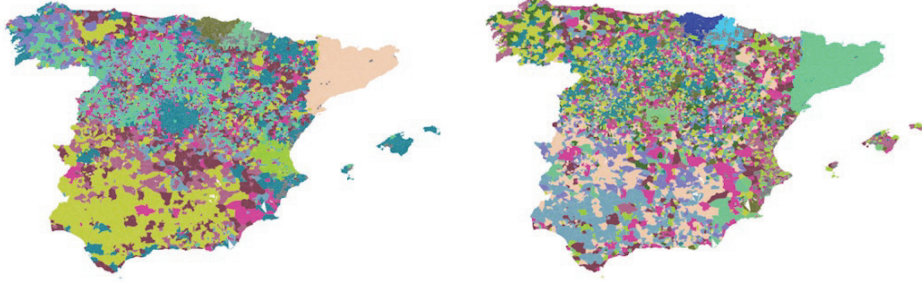


FIGURE 6: First-level communities obtained via community detection algorithm Infomap, performed on the thresholded, undirected, and weighted network based on the similarity matrix \mathbf{S} (see the text). Here we color nodes belonging to the same community in the same color. There is an appreciable matching between functional network communities in the network of municipalities and actual Spanish autonomous communities, although this correspondence is much more evident for the 2015 round, where bipartisanship breakdown was at its zenith.

accordingly mismanaged or fraudulent manipulation of vote counts would induce a statistically significant deviation in the distribution of the second leading digit, detected by a simple Pearson χ^2 goodness-of-fit test. Mebane applied this so-called 2BL test to assess the cases of Florida 2004 and Mexico 2006, and other authors have subsequently applied this in many other occasions (see [16] and references therein). In this case

the theoretical distribution takes a more convoluted shape than (4), namely,

$$P_2(d) = \sum_{k=1}^9 \log_{10} \left(1 + \frac{1}{10k+d} \right), \quad (5)$$

and a good numerical approximation [15, 16] is given by

$$P_2(d) \approx (0.11968, 0.11389, 0.10882, 0.10433, 0.10031, 0.09668, 0.09337, 0.09035, 0.08757, 0.08500). \quad (6)$$

We start by exploring 1BL and 2BL tests applied to vote count statistics nationally using the fine-grained data given by splitting vote counts at the level of municipalities (with over 8000 samples, vote counts ranging in about five orders of magnitude). With sociopolitical impact in mind, from now on we focus on the vote statistics to the main, national-wide parties PP, PSOE, Cs, and Podemos.

Results for the 1BL are shown in Figures 7(a) and 7(c) ((a, c) depicts results for the 2015 elections while (b, d) does the same for the 2016 case). As expected the distributions seem to be close to Benford's law for all political parties, at least visually, and there are no obvious differences between 2015 and 2016. To have a better quantitative understanding, we have made use of two statistics: (i) the classical Pearson's χ^2 and (ii) the mean absolute deviation (MAD) test as proposed by Nigrini [3]. In both cases the null hypothesis H_0 is that data conform to Benford's law. The former statistic reads

$$\chi^2 = N \sum_{d=m}^9 \frac{[P_{\text{obs}}(d) - P_{\text{th}}(d)]^2}{P_{\text{th}}(d)}, \quad (7)$$

where $P_{\text{th}}(d)$ and $P_{\text{obs}}(d)$ are the theoretical and observed relative frequencies of each digit and $m = 1$ for 1BL and $m = 0$ for 2BL. This statistic has 8 degrees of freedom for 1BL and 9 for 2BL (as in this latter case the digit zero has to be incorporated as a candidate) and is to be compared to certain critical values, such that if $\chi^2 > \chi_{n,a}^2$ then H_0 is rejected with the selected level of confidence level a . For $n = 8$ degrees of freedom, the critical values at the 95% and 99% are 15.507 and

20.090, respectively, whereas, for $n = 9$ degrees of freedom, the critical values at the 95% and 99% are 16.919 and 21.666, respectively.

The mean absolute deviation is defined as

$$\text{MAD} = \frac{1}{10-m} \sum_{d=m}^9 |P_{\text{obs}}(d) - P_{\text{th}}(d)|, \quad (8)$$

where m is the initial digit (1 for 1BL, 0 for 2BL). Whereas this statistic lacks clear cut-off values, Nigrini provides the following rule of thumb for 1BL: MAD between 0 and 0.004 implies close conformity; from 0.004 to 0.008 acceptable conformity; from 0.008 to 0.012 marginally acceptable conformity; and, finally, greater than 0.012 nonconformity. To the best of our knowledge, the critical values for MAD have not yet been established for 2BL so all over this work we will assume the same ones as for 1BL.

Results for 1BL can be found in Table 1. We conclude that, for the Pearson χ^2 test, H_0 cannot be rejected with sufficiently high confidence in three out of the four main political parties but the χ^2 result for PP is consistently large and suggests rejection of the null hypothesis with a confidence of 99%. These results are in contrast with those found using the MAD statistic, where according to Nigrini all political parties conform to Benford's law (PP only showing acceptable conformity and the rest showing close conformity).

The results on 2BL are shown in Figures 7(b) and 7(d) and test statistics are summarized again in Table 1. These suggest an overall conformance to the second digit law, with

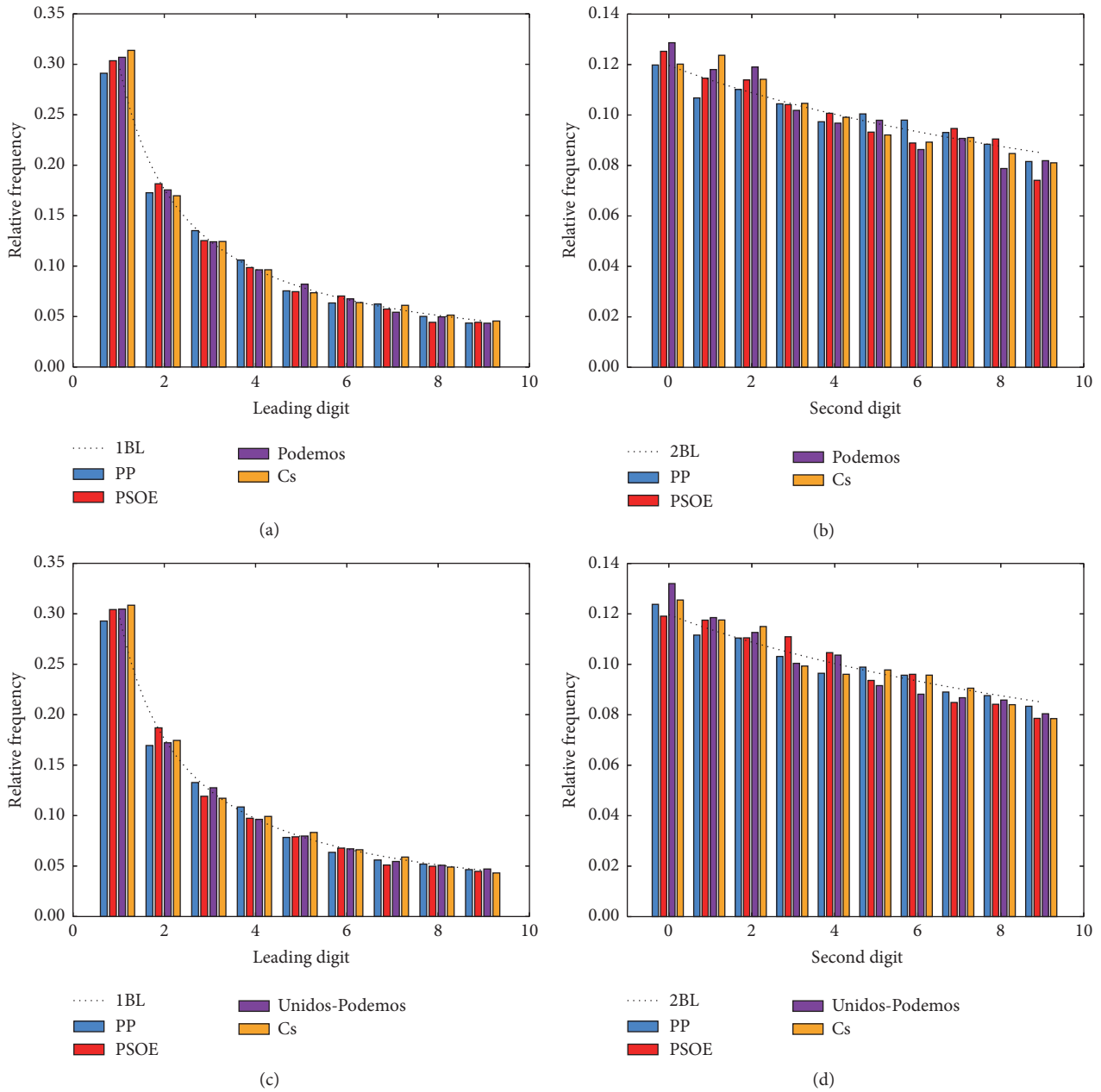


FIGURE 7: Histograms of relative frequencies for the first (a, c) and second (b, d) significant digits of the four most important political parties vote counts over municipalities (more than 8000 in each case) for the 2015 (a, b) and 2016 (c, d) elections.

exception flagging nonconformance raised by χ^2 that rejects H_0 at 95% for Podemos (2015), Unidos-Podemos (2016), and PSOE (2015).

4.1.1. Individual Analysis at the Precincts Level. In order to give a closer look to the vote count distributions we now explore the statistics taking place at each separate precinct. At this point we need to recall that among other criteria Benford's law is expected to emerge in datasets where data range several orders of magnitude. This hypothesis was fulfilled at the national scale as the population of municipalities ranges several orders of magnitude ($\mathcal{O}(1)$ – $\mathcal{O}(10^5)$, see Table 2) if we

consider all of them. However this is not straightforward at the precinct level, where the number of municipalities is highly heterogeneous from precinct to precinct. In Figure 8 we have checked that the number of orders of magnitude that vote counts span at the precinct level is indeed linearly correlated with the number of municipalities the precinct contains ($R^2 = 0.47$). This means that the larger the number of municipalities considered in a single analysis is, the more we should expect data to conform to Benford's law.

That being said, for each and every precinct in Spain, we proceed to extract the frequencies of the first and second significant digits found for all the municipalities inside that

TABLE 1: Statistical tests of conformance to Benford’s law for the first (1BL) and second (2BL) significant digit distribution for the vote counts of each political party (at the level of municipalities), along with χ^2 and MAD statistics. In *italic* we highlight the datasets where the null hypothesis can be rejected with 95% confidence but not with 99% and in **bold** cases for which where the null hypothesis can be rejected with more than 99% confidence according to χ^2 . On the basis of MAD statistic the null hypothesis of conformance to Benford’s laws cannot be rejected for any case.

Year	Political party	Number of observations	χ^2 1BL	MAD 1BL	χ^2 2BL	MAD 2BL
2015	PP	8182	23.079	0.0052	8.737	0.0027
2016	PP	8186	21.408	0.0046	4.142	0.0021
2015	PSOE	8135	13.486	0.0030	<i>17.648</i>	0.0038
2016	PSOE	8121	15.040	0.0033	13.065	0.0038
2015	Podemos & Co.	7927	4.845	0.0020	<i>21.329</i>	0.0050
2016	Unidos Podemos	8056	3.537	0.0019	<i>18.314</i>	0.0048
2015	C’s	8037	11.933	0.0036	10.934	0.0034
2016	C’s	8001	9.671	0.0033	10.951	0.0039

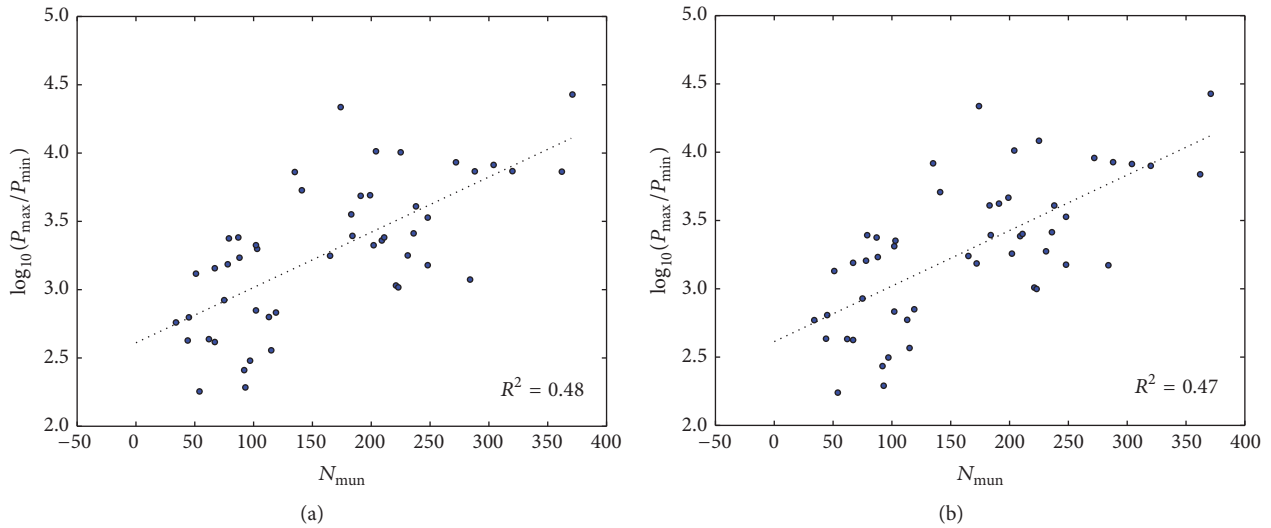


FIGURE 8: Scatter plot of the number of orders of magnitude spanned by the voting populations of a precinct as a function of the number of municipalities in each precinct, for 2015 (a) and 2016 (b). We find a positive correlation with $R^2 \approx 0.47$ suggesting that the “size” of a precinct in terms of the number of municipalities explains 47% of the variation in the support (in terms of orders of magnitude) of the number of votes (the larger the number of municipalities is, the more likely the number of votes takes values from a larger number of orders of magnitude). The coefficient of 0.004 indicates that, on average, when we move from a precinct with x municipalities to one with $x + 100$, the ratio of the biggest population to the smallest is 2.5 times bigger.

precinct and make a goodness-of-fit test between these empirical distributions and 1BL and 2BL using both χ^2 and MAD statistics. Results on 1BL are summarized for the case of χ^2 in Figures 9(a) and 9(c) and in Figure 10, finding an overall good conformance to 1BL at the precinct level. Conversely, MAD statistics (Figure 11) say just the opposite, suggesting systematic nonconformance. As for the 2BL test, there exists a strong deviation from the expected distribution (Figures 9(b) and 9(d) for χ^2 and Figure 11 for MAD), and both statistics consistently reject the null hypothesis of conformance to Benford’s law for all political parties.

Now, note that at each individual precinct we expect statistics to be a priori poorer than at the national scale, as the average number of municipalities per precinct is of the order of $\mathcal{O}(10^2)$ (see Table 2 for details), that is, one order of magnitude smaller. As MAD does not include any

correction term that depends on the sample set, one should therefore take the results associated with MAD with a pinch of salt. This is not necessarily the case for the χ^2 as this latter statistic takes into account in its definition the number of samples. In any case, in order to assess whether the strong nonconformance to 2BL at this level of aggregation is just due to finite size effects we explore the dependence of both χ^2 and MAD results on the precinct’s size. Accordingly, in Figure 12 we plot for each precinct its χ^2 and MAD result as a function of the number of municipalities present in that precinct. As expected, we find that MAD suffers from finite size effects and is over conservative for small sample sizes; however, this effect is rather weak and not enough to explain the systematic nonconformance to 2BL. In the case of the χ^2 statistic we observe quite the opposite effect: the larger the number of municipalities in a given precinct, the

TABLE 2: List of precincts with their number of different municipalities and the voting population ranges (by voting population we mean the number of possible voters). The largest cities, such as Madrid, Barcelona, Bilbao, Sevilla, Valencia, Zaragoza and Malaga, have been subsequently divided into electoral districts and we have treated these latter districts as municipalities.

Precinct	N_{mun}	Range 2015	Range 2016
Cadiz	44	383–162564	376–162111
Tarragona	184	36–89136	36–89031
A-Coruna	93	1020–196251	1007–196492
Zaragoza	304	12–98267	12–98399
Valencia-Valencia	284	37–43858	39–58015
Leon	211	42–101272	40–100862
Avila	248	13–43810	13–43759
Gipuzkoa	88	85–145679	85–145167
Granada	172	57–182735	119–182450
La-Rioja	174	5–108493	5–108755
Lugo	67	186–76951	183–77209
Castellon-Castello	135	16–116252	14–116049
Jaen	97	297–89693	285–89487
Cordoba	75	305–255629	301–255476
Barcelona	320	25–184321	23–182889
Araba-Alava	51	140–183368	136–183559
Valladolid	225	24–243129	20–242609
Teruel	236	10–25834	10–25959
Ourense	92	331–85240	314–85329
Palencia	191	13–63236	15–63072
Navarra	272	17–145462	16–145189
Asturias	78	146–223974	139–223268
Huelva	79	47–111520	45–111093
Pontevedra	62	535–232242	542–232465
Soria	183	8–28459	7–28540
Madrid	199	36–176867	38–176527
Sevilla	115	272–97848	266–98048
Huesca	202	18–38062	21–37988
Illes-Balears	67	192–275448	178–275883
Lleida	231	51–90807	48–90289
Cantabria	102	64–135418	66–135258
Murcia	45	492–308510	482–309387
Malaga	113	133–83852	142–84163
Ciudad-Real	102	81–57081	84–57248
Cuenca	238	10–40719	10–40722
Caceres	223	70–72783	75–74773
Segovia	209	17–38948	16–38867
Guadalajara	288	8–58795	7–59179
Girona	221	59–63288	62–63305
Salamanca	362	16–116942	17–117091
Almeria	103	70–139271	62–139412
Bizkaia	119	116–78875	111–78573
Toledo	204	6–61813	6–61731
Santa-Cruz-de-Tenerife	54	887–159534	919–159695
Albacete	87	54–130156	55–130572
Alicante-Alacant	141	44–234975	46–234691
Las-Palmas	34	508–292289	496–292504

TABLE 2: Continued.

Precinct	N_{mun}	Range 2015	Range 2016
Badajoz	165	63–111575	66–114755
Zamora	248	34–51358	34–51049
Burgos	371	5–134171	5–133923

TABLE 3: χ^2 values of conformance to 1BL and 2BL for each political party extracted from the analysis performed when we aggregate votes at the precinct level.

Year	Political party	test	χ^2	MAD
2015	PP	1BL	12.71	0.0455
2016	PP	1BL	7.13	0.0329
2015	PP	2BL	7.79	0.0325
2016	PP	2BL	7.43	0.0333
2015	PSOE	1BL	3.78	0.02628
2016	PSOE	1BL	2.90	0.0220
2015	PSOE	2BL	20.90	0.0541
2016	PSOE	2BL	4.15	0.0222
2015	Podemos & Co.	1BL	4.45	0.02638
2016	Unidos Podemos	1BL	7.09	0.0344
2015	Podemos & Co.	2BL	3.29	0.0168
2016	Unidos Podemos	2BL	8.46	0.0359
2015	C's	1BL	4.81	0.02433
2016	C's	1BL	5.24	0.0270
2015	C's	2BL	15.39	0.0474
2016	C's	2BL	8.86	0.0289

more likely the null hypothesis to be rejected. An equivalent size dependence analysis for the 1BL test is reported in Figure 13.

4.1.2. Aggregate Analysis at the Precincts Level. To round off our analysis with a third level of aggregation, we explore conformance to 1BL and 2BL when vote counts are aggregated per precinct. In this case, we only have 52 samples (52 precincts) so we expect the distributions to be more noisy. From the previous analysis, we learned that MAD suffers from finite size effects so we expect MAD to be more conservative than χ^2 at this level of aggregation. In Figure 14 we show the results for 2016 and we refer the readers to Figure 15 to find analogous results for 2015, which do not show substantial differences at the qualitative level. As expected the distributions show larger fluctuations and, in absolute terms, deviate more from the theoretical laws (depicted in dotted lines). In Table 3 we depict χ^2 and MAD statistics, which, again as expected, show inconsistent results: while χ^2 systematically cannot be rejected with above 95% confidence level, in turn MAD systematically suggests nonconformity. We conclude that this level of aggregation is less informative than previous ones.

4.2. Cooccurrence Heat Maps. Our second analysis is inspired by a recent study [5] that explore the cooccurring statistics

of vote and turnout numbers and the associated double mechanism of incremental and extreme fraud by plotting two-dimensional histograms (heat maps) reporting, for a given political party, the percentage of vote (vote share) it got as compared to the percentage of participation. According to Klimek and coauthors, incremental fraud occurs when, with a given rate, ballots for one party are added to the urn and/or votes for other parties are taken away, and this mechanism is revealed when the histograms smear out towards the top-right corner of the histograms. On the other hand, extreme fraud (which corresponds to reporting close to complete turnout and almost all votes for a single party) emerges when the distribution transitions from unimodal to bimodal and one of the modes corresponding to a cluster that concentrates close to that corner of 100% participation (complete turnout) and very large vote percentage. They applied these statistical principles to several national elections, concluding that in the cases of Russia and Uganda fraudulent manipulation was the most likely underlying mechanism. In Figure 16 we plot such heat maps for the 2016 case for all four political parties. Data for PSOE, Unidos-Podemos, and Cs do not show any sign of fraudulent manipulation. In the case of PP results are less clear, as there indeed exists a (rather weak) tendency of the data to smear out towards the top-right corner (results for 2015 are very similar and have been reported in Figure 17). We do not find any sign of systematic extreme fraud, although it is worth stating that we have found a small subset of municipalities where just one party received 100% of the vote share (see Table 4). Without exception, this is the popular party (PP), something that is in principle suspicious. Nevertheless a closer inspection reveals that these municipalities are extremely small and thus consensus in one political option cannot be ruled out statistically.

A further interesting peculiarity for the case of the conservative party PP is the existence of two clusters of municipalities (bimodal distribution) that gathers two different voting strategies: one relatively small, located at small vote share and the other one at high vote share, which is more spread out (we do not observe bimodality for the rest of political parties). We have labeled municipalities according to which cluster they belong (assigning a brown label for the larger cluster and a turquoise label for the smaller one) and plotted them in Figure 18. Just by visual inspection we can appreciate that the category linked with the smaller cluster is mainly formed by Catalonia and the Basque Country (regions with proindependence aspirations and a strong nationalist tradition), something that was recently pointed out independently [17], and some further municipalities in regions that have been considered PSOE strongholds historically.

TABLE 4: List of municipalities where a single party receives 100% of the vote share.

Year	Municipality	Population	Turnout	Party receiving 100% of the vote share
2015	Castilnuevo (Guadalajara)	8	88%	PP
2015	Valdemadera (La Rioja)	7	100%	PP
2016	Castilnuevo (Guadalajara)	7	100%	PP
2016	Rebollosa de Jadraque (Guadalajara)	10	90%	PP
2016	Congostrina (Guadalajara)	16	62%	PP
2016	La Vid de Bureba (Burgos)	16	65%	PP
2016	Portillo de Soria (Soria)	16	75%	PP
2016	Valdemadera (La Rioja)	8	88%	PP

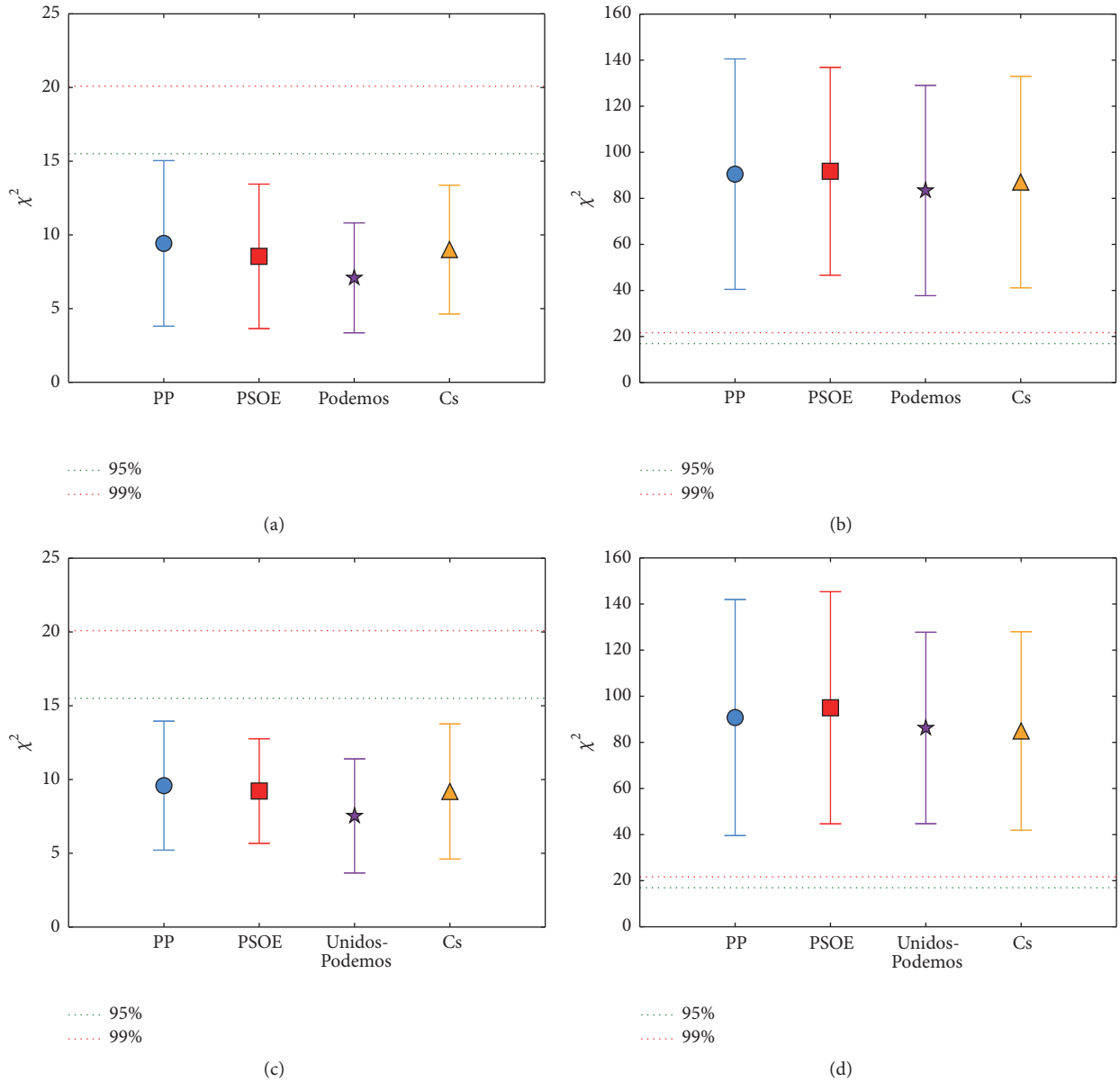


FIGURE 9: Summary of Pearson χ^2 goodness of fit to 1BL (a, c) and 2BL (b, d) for 2015 (a, b) and 2016 (c, d) extracted from analysis of each individual precinct (each precinct contains a different number of municipalities and shows a precise distribution and an associated χ^2 , so here we plot the mean \pm standard deviation over all Spanish precincts (excluding Ceuta and Melilla, precincts with a single municipality)) for the main parties. In every case, the critical values for rejection at the 95 and 99% confidence level are shown. Interestingly, in the case of 1BL for a large majority, we accept conformance to Benford’s law, whereas in the case of 2BL for a large majority the null hypothesis is rejected. Results based on MAD suggest that neither 1BL nor 2BL is accepted (Figure 11).

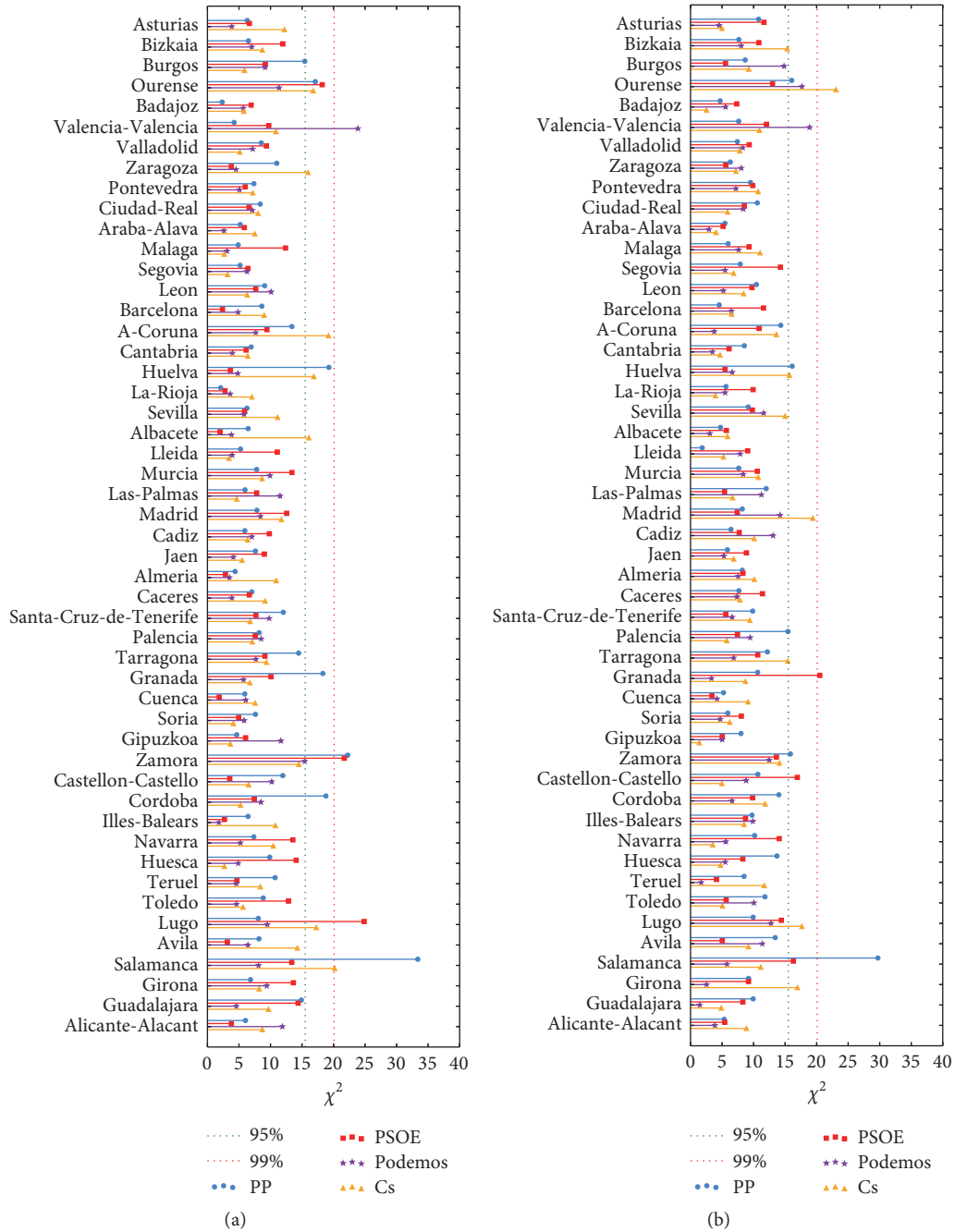


FIGURE 10: χ^2 values of the goodness of fit to 1BL for 2015 (a) and 2016 (b) at the aggregation level of precincts. In every case the critical values for rejection at the 95 and 99% confidence level are shown. For a large majority we accept conformance to Benford's law. Note that results are inconsistent with the hypothesis test based on MAD as reported in Figure 19.

5. Discussion

In this work, we have studied the statistical properties of vote counts in the Spanish national elections that took place in December 2015 and June 2016, focusing on three separate questions: (i) breakdown of bipartisanship, (ii) region-to-region similarity in vote percentage, and (iii) election forensics for fraud detection.

On relation to (i), our results highlight that the bipartisanship system has suffered a clear breakdown in 2015, at least in regions associated with a more widespread cosmopolite society. Such breakdown consolidates over time but does not increase, probably due to the risk aversion of not finding workable majorities in the second election round and even evidences a subtle decrease as captured by the diversity index N_{eff} . Bipartisanship breakdown is actually a quite

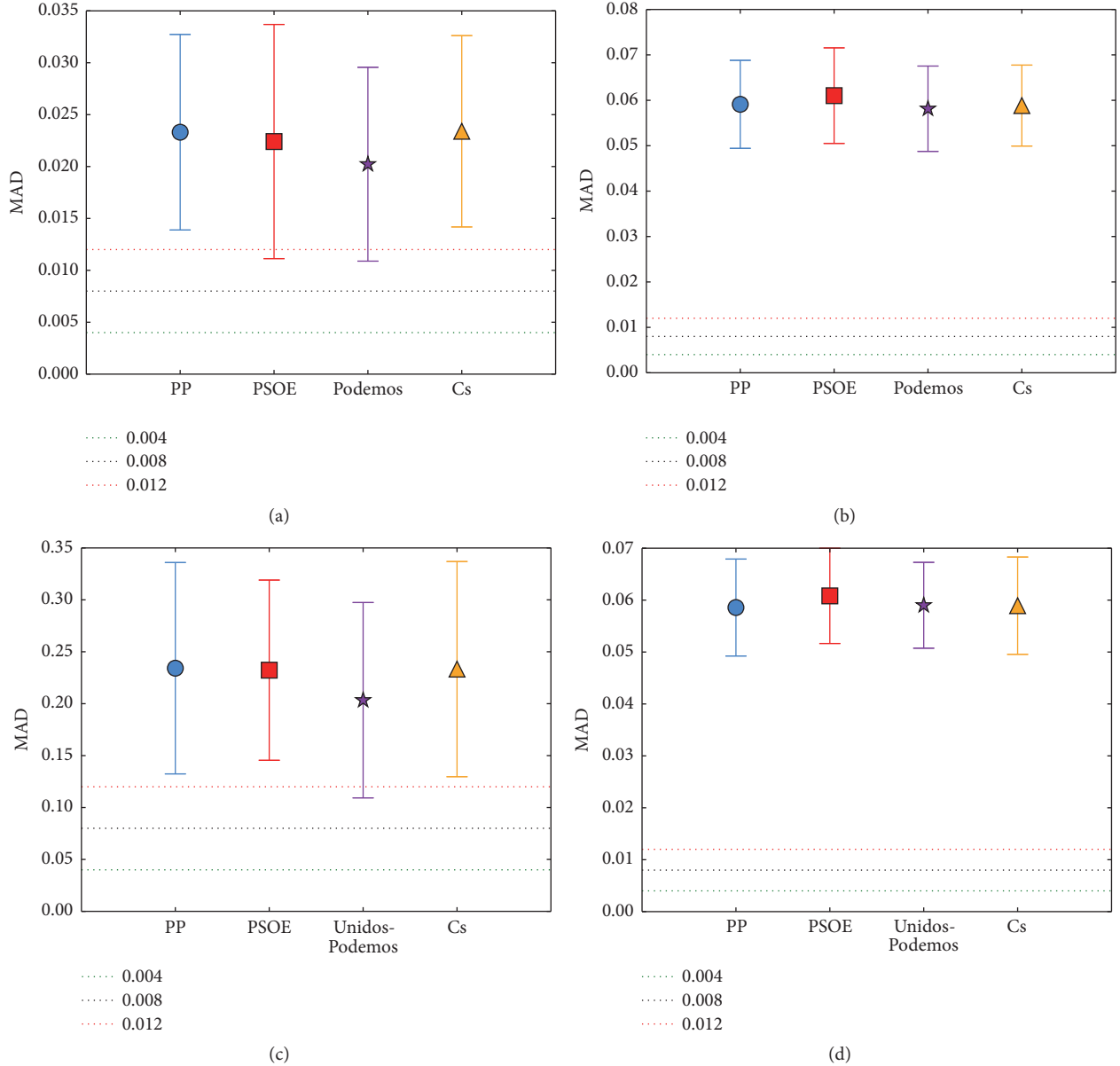


FIGURE 11: Summary of MAD goodness of fit to 1BL (a, c) and 2BL (b, d) for 2015 (a, b) and 2016 (c, d), performed individually at each precinct (each precinct shows a precise distribution and an associated MAD, so here we plot the mean \pm standard deviation over all Spanish precincts, excluding Ceuta and Melilla, precincts with a single municipality). In every case, the critical values for rejection at the 95 and 99% confidence level are shown. Interestingly, in the case of 1BL for a large majority, we accept conformance to Benford's law, whereas, in the case of 2BL for a large majority, the null hypothesis is rejected. All these results are consistent with the hypothesis test based on MAD.

complex phenomenon with a high degree of heterogeneity at a regional level, probably due to regional political particularities.

Second, on relation to (ii), we have constructed a functional network of municipalities via cosine similarity of voting profiles. Interestingly, there is a very good matching between network communities which emerge by a community detection algorithm on the functional network and the actual Spanish autonomous communities. In particular, a classification of autonomous communities emerges naturally: we find some autonomous communities whose functional

network community counterpart is more cohesive and stable over time (e.g., Catalonia, Basque Country, Madrid, and Navarra), some whose counterpart is only well-defined in 2015 when bipartisanship breakdown was more acute (e.g., Murcia and Valencia), and the rest where there is no clear matching. Beyond the probably amplified role played by regionalist parties in 2015, we do not have clear sociopolitical explanations for such emergent classification and we leave this as an open problem. Other aspects left for future work include network pruning using different criteria to the ones applied in this work, such as using a fixed similarity threshold.

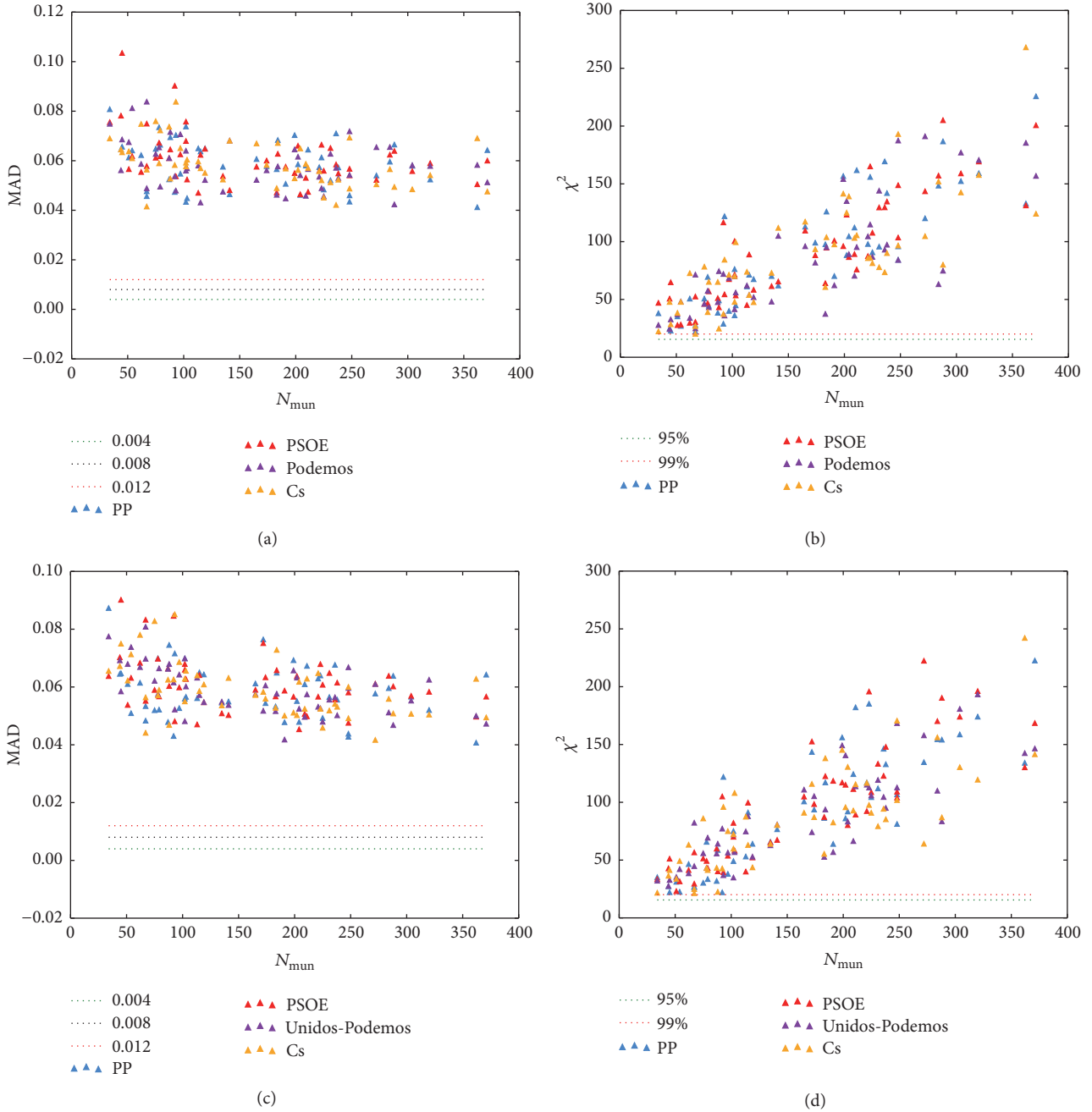


FIGURE 12: Scatter plot of the MAD (a, c) and χ^2 (b, d) statistics extracted from the 2BL test of each precinct as a function of the number of municipalities in each precinct (2015 results are shown in (a, b) and 2016 ones are shown in (c, d), with no obvious differences). In the case of MAD, we find a weak negative correlation as expected, but this correlation is not enough to explain the systematic nonconformance to 2BL. In the case of χ^2 , the effect is quite the opposite, and nonconformance is stronger as the size of the precinct increases, thereby suggesting that nonconformance to 2BL at this level of aggregation is a genuine result and not a spurious effect of finite size statistics.

On relation to fraud detection, for the 2016 elections, the unusually high discrepancy found between electoral surveys preceding and on the day of the elections (26th June) and the actual electoral results have been a source of debate and controversy in Spanish media. To the best of our knowledge, this work is among the first systematic analysis of its kind for Spanish elections (see however [17, 18]). The first and general

conclusion on relation to question (iii) we have extracted is that the voting distributions do not show any systematic and significant change between the 2015 and the 2016 elections, as all statistical results are qualitatively identical. This is in line with the original analysts thesis that were discussed soon after it was learned that Spain had to go into a second election given the inability of the parliament to find a suitable coalition, but

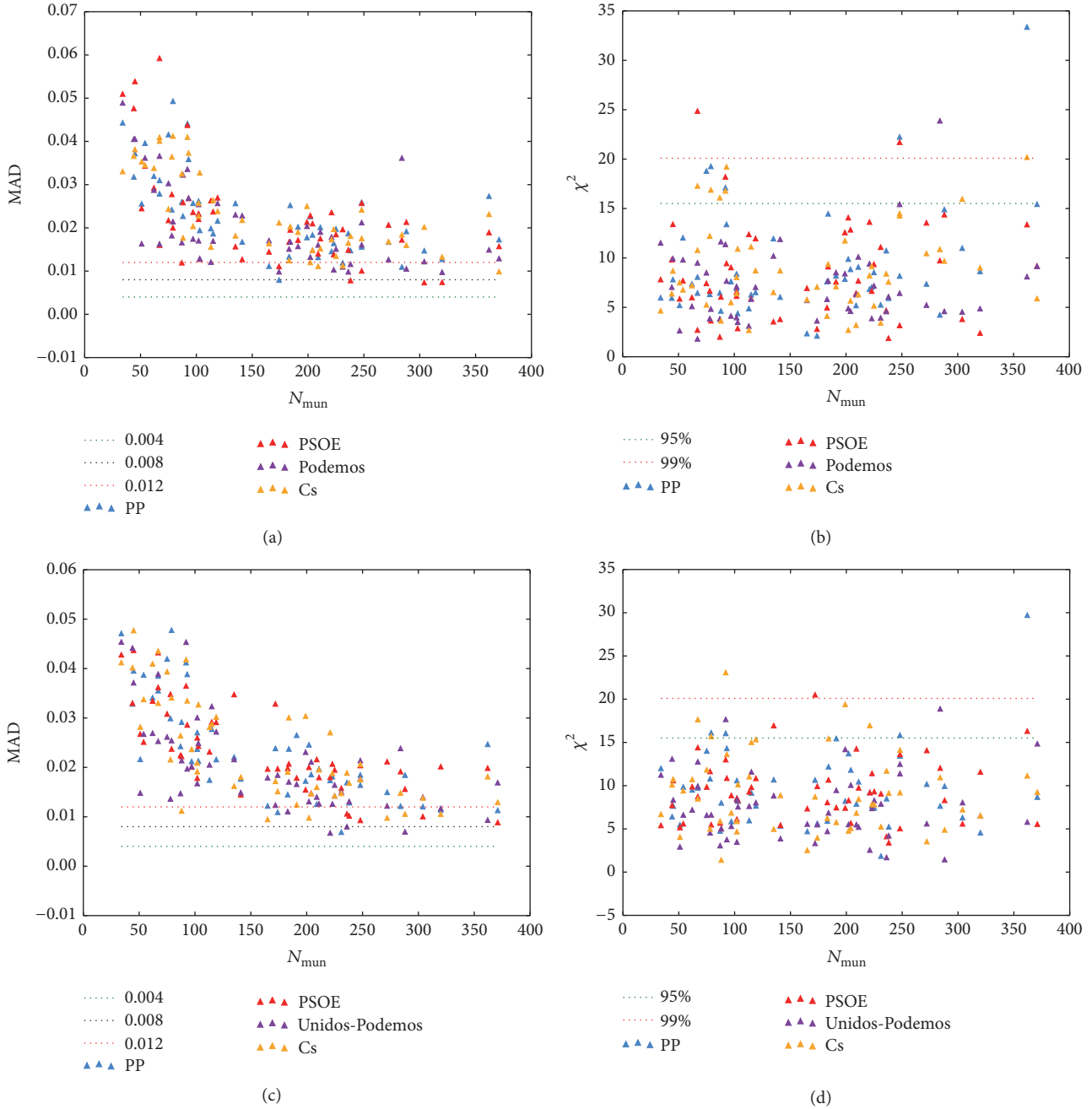


FIGURE 13: Scatter plot of the MAD (a, c) and χ^2 (b, d) statistics extracted from the IBL test of each precinct as a function of the number of municipalities in each precinct for years 2015 (a, b) and 2016 (c, d). In the case of MAD, we find a negative correlation as expected, but this correlation is not enough to explain the systematic nonconformance to IBL. In the case of χ^2 there is no perceivable size effect.

at odds with most of the polls and surveys of vote intention which were predicting a much different scenario as 26th June approached.

The first analysis is based on the hypothesis that, under clean conditions, vote count data should conform to Benford's law. At the national scale we have found a general good qualitative and quantitative conformance to Benford's law for the first (1BL) and second (2BL) digits, with small deviations only occurring for IBL in the conservative party, where the

null hypothesis can be rejected at 99% confidence in both years according to the standard Pearson χ^2 hypothesis test, a result which is not confirmed using an alternative test (mean absolute deviation) proposed by Nigrini. For 2BL only χ^2 flags up some concerns at the 95% confidence level for Podemos/Unidos-Podemos, but the null hypothesis cannot be rejected at 99% and, again in this case, MAD statistic is less conservative and accepts the null hypothesis for every party.

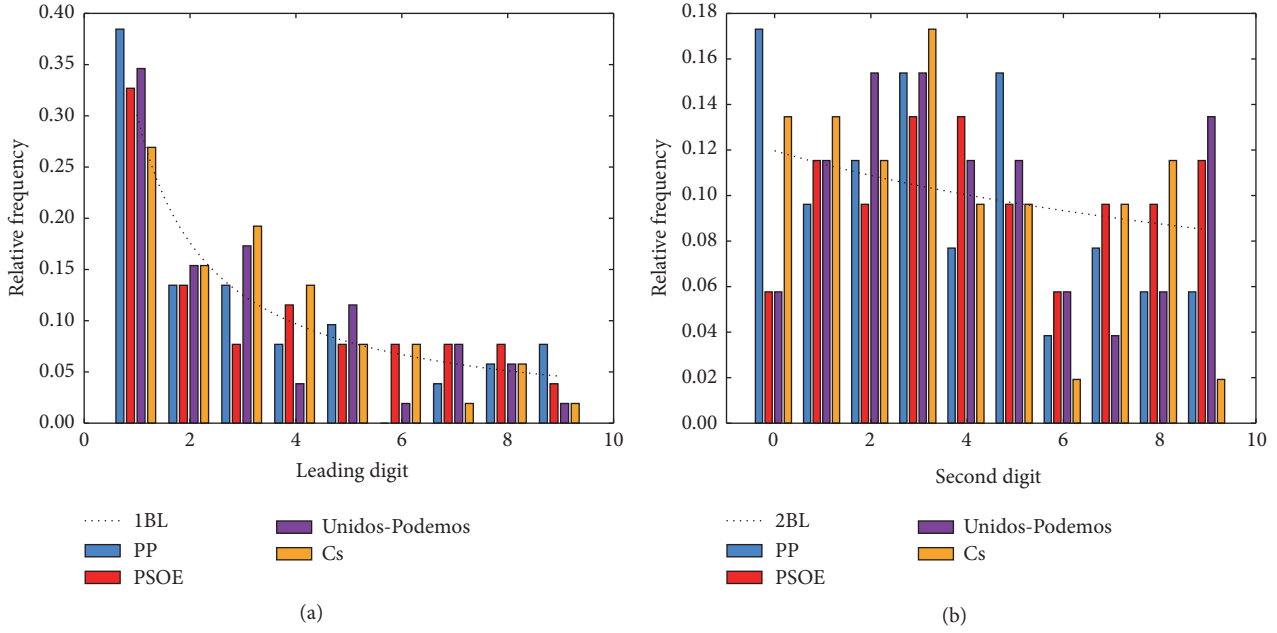


FIGURE 14: Histograms of relative frequencies for the first (a) and second (b) significant digits for the main political parties vote counts aggregated over precincts, for the case of 2016 (2015 is shown in Figure 15).

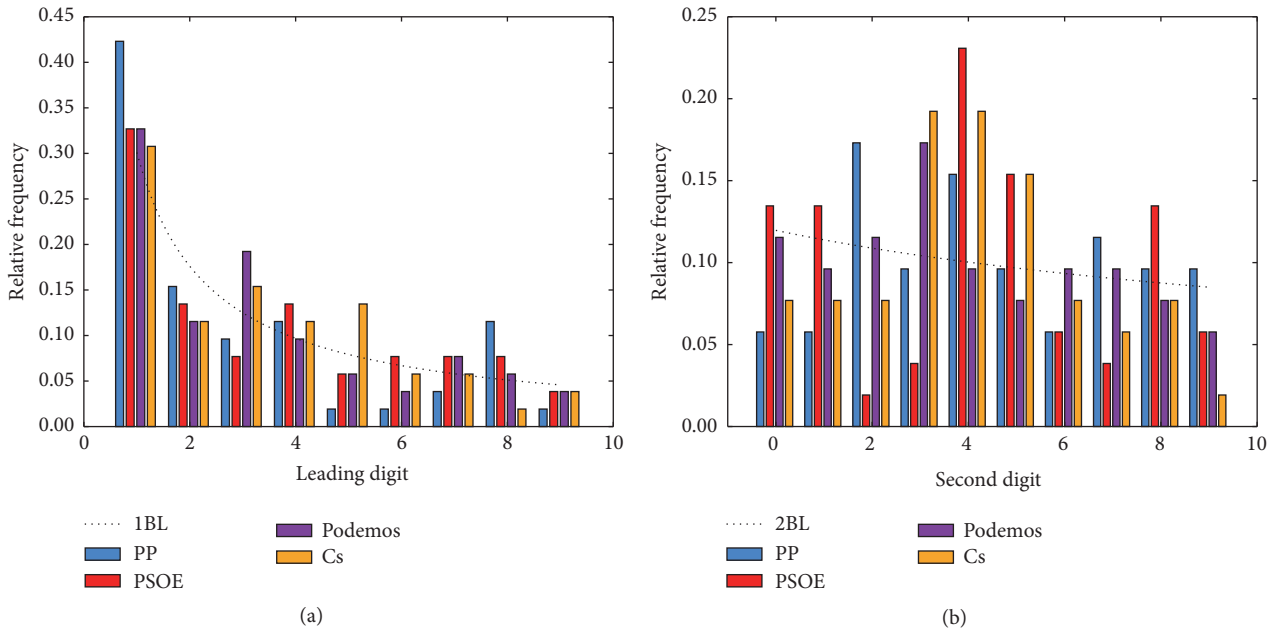


FIGURE 15: Histograms of relative frequencies for the first (a) and second (b) significant digits for the main political parties vote counts aggregated over precincts, for the case of 2015.

If we change the resolution and explore results for each individual precinct, results show a completely different story: conformance to 1BL is accepted according to χ^2 but systematically rejected according to MAD, and conformance to 2BL is consistently rejected according to both χ^2 and MAD statistics for every precinct and every political party. We have also shown that these are genuine results that cannot be associated with a lack of statistics. Finally, by aggregating

vote counts per precinct and analyzing conformance to 1BL and 2BL at this level of aggregation, we obtain inconsistent and therefore inconclusive results, as χ^2 cannot reject the null hypothesis above 95% confidence level systematically but conversely MAD suggests systematic nonconformance. This lack of consistency raises the question about what level of aggregation might be better suited for BL-type analysis and which statistic is more reliable when assessing

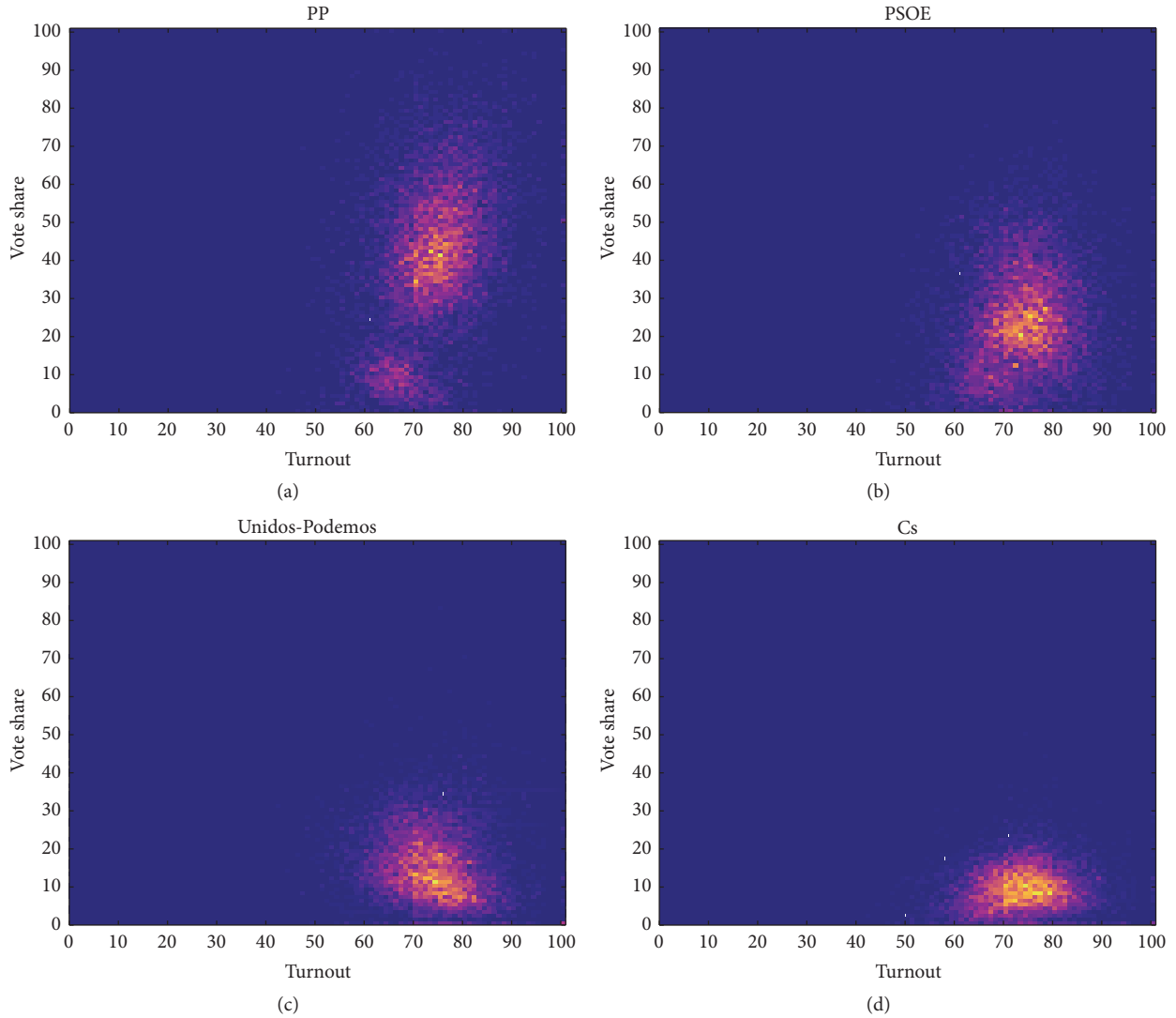


FIGURE 16: Heat maps plotting the percentage (in color scale) of municipalities where a given political party has received a certain percentage of votes, as a function of the relative participation. These are results from the 2016 elections; the 2015 case is reported in Figure 17. According to Klimek et al. [5], a smear out of the cluster towards the top-right corner of the heat map is a sign of incremental fraud, whereas extreme fraud would occur for bimodal distributions where a cluster emerges at the top-right corner.

the goodness-of-fit issues that certainly deserve further investigation.

Given the somewhat mixed results and acknowledging that the applicability of Benford's law tests to election forensic is not completely free from controversy [19, 20], as a complementary analysis we further explored the correlations between percentage of participation and percentage of votes for each municipality, plotting two-dimensional histograms to detect the presence of so-called incremental and/or extreme fraud as described by Klimek et al. [5]. Our results suggest that the results for PSOE, Unidos-Podemos, and Cs are apparently free from these mechanisms whereas in the case of PP we find a weak evidence of cluster smearing out similarly to what Klimek et al. refer to incremental fraud, an evidence which needs to be studied in more detail. The heat map of the conservative party also shows two

clusters instead of a single one, hence the bimodality in the vote share tendency: there exist two different groups of municipalities, including a small one where the tendency is to give a small vote share to PP and a larger one where the vote share takes larger values. Interestingly, according to a spatial analysis, we have been able to confirm that the low vote share cluster typically corresponds to regions which are considered nationalist (Catalonia and Basque Country) where the strength of regional options outperforms those that prevail at a nationwide scale.

All in all, these results suggest that further investigations and enquiries should be conducted in order to confirm and clarify the presence or absence of some of these apparent irregularities, to elucidate their source and quantify their impact in election results. In this respect, systematic comparative studies with historical bipartisanship Spanish data

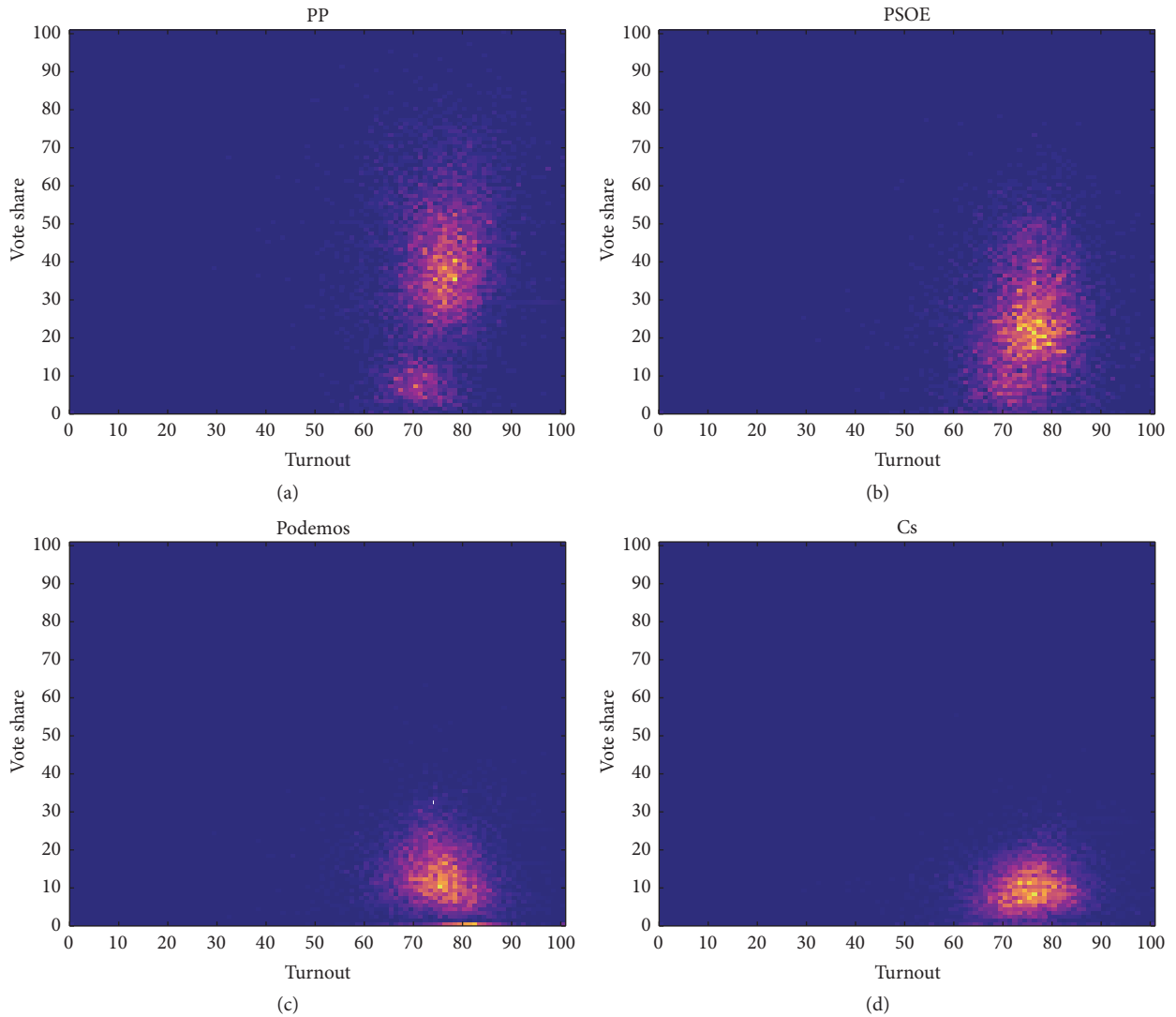


FIGURE 17: Heat maps plotting the percentage (in color scale) of municipalities where a given political party has received a certain percentage of votes, as a function of the relative participation. These are results associated with December 2015 elections. According to Klimek et al. [5], a smear out of the cluster towards the top-right corner of the heat map is a sign of incremental fraud, whereas extreme fraud would occur for bimodal distributions where a cluster emerges at the top-right corner.

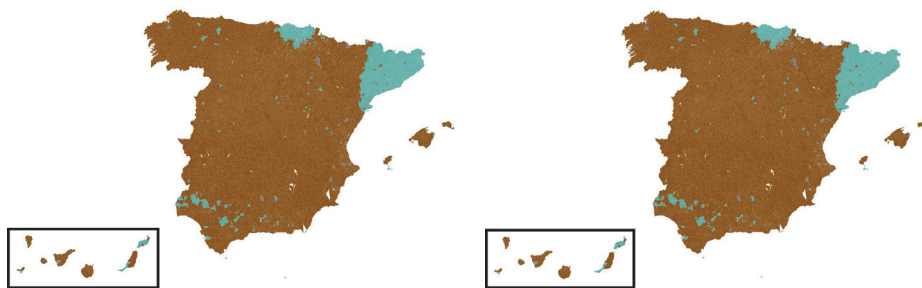


FIGURE 18: Focusing on the bimodal distribution for the conservative party that emerges in the heat map of Figure 16, here we show in a spatial map of Spain where we assign a brown color to those municipalities that belong to the larger cluster (high vote share) and a turquoise color to those that belong to the smaller cluster (low vote share). We find that the low vote share clusters are predominantly linked with Catalonia and the Basque Country, the two areas of Spain with some proindependence aspirations. No obvious change is perceived between 2015 and 2016.

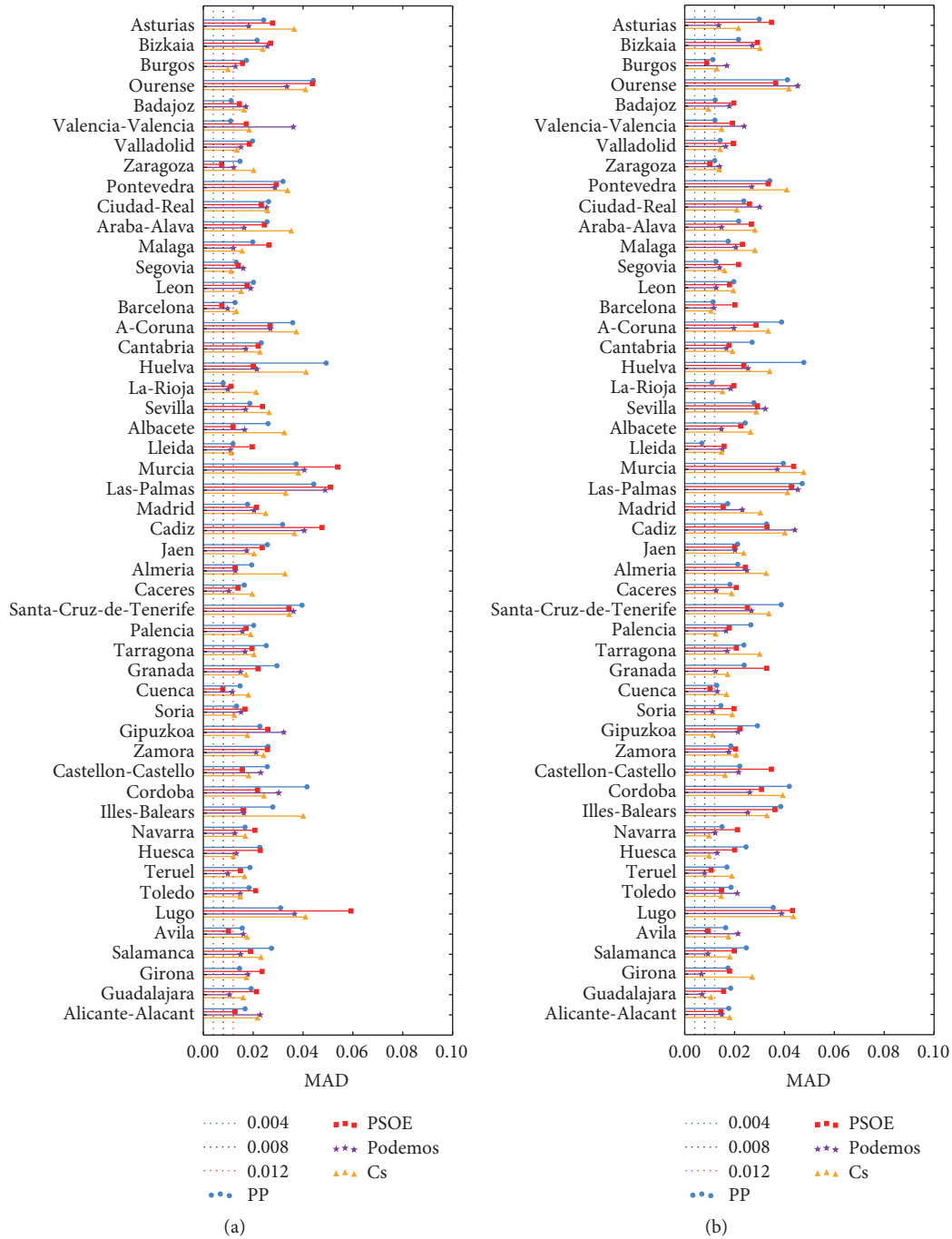


FIGURE 19: MAD values of the goodness of fit to IBL for 2015 (a) and 2016 (b) at the aggregation level of precincts. In every case, the critical values for rejection at the 95 and 99% confidence level are shown. For a large majority, we reject conformance to Benford's law at the precinct level according to MAD, this result being inconsistent with the one found for Pearson's χ^2 statistic.

and analogous data (analysis at different levels of aggregation) from other similar democratic countries are needed.

Appendix

In this appendix we depict several additional figures and tables that complement the main study (see the main text for

references to each of these figures). See Figures 3, 8, 10, 11, 13, 15, 17, 19, 20, and 21 and Table 2.

Additional Points

Availability of Data and Materials. The datasets supporting the conclusions of this article are available under request.

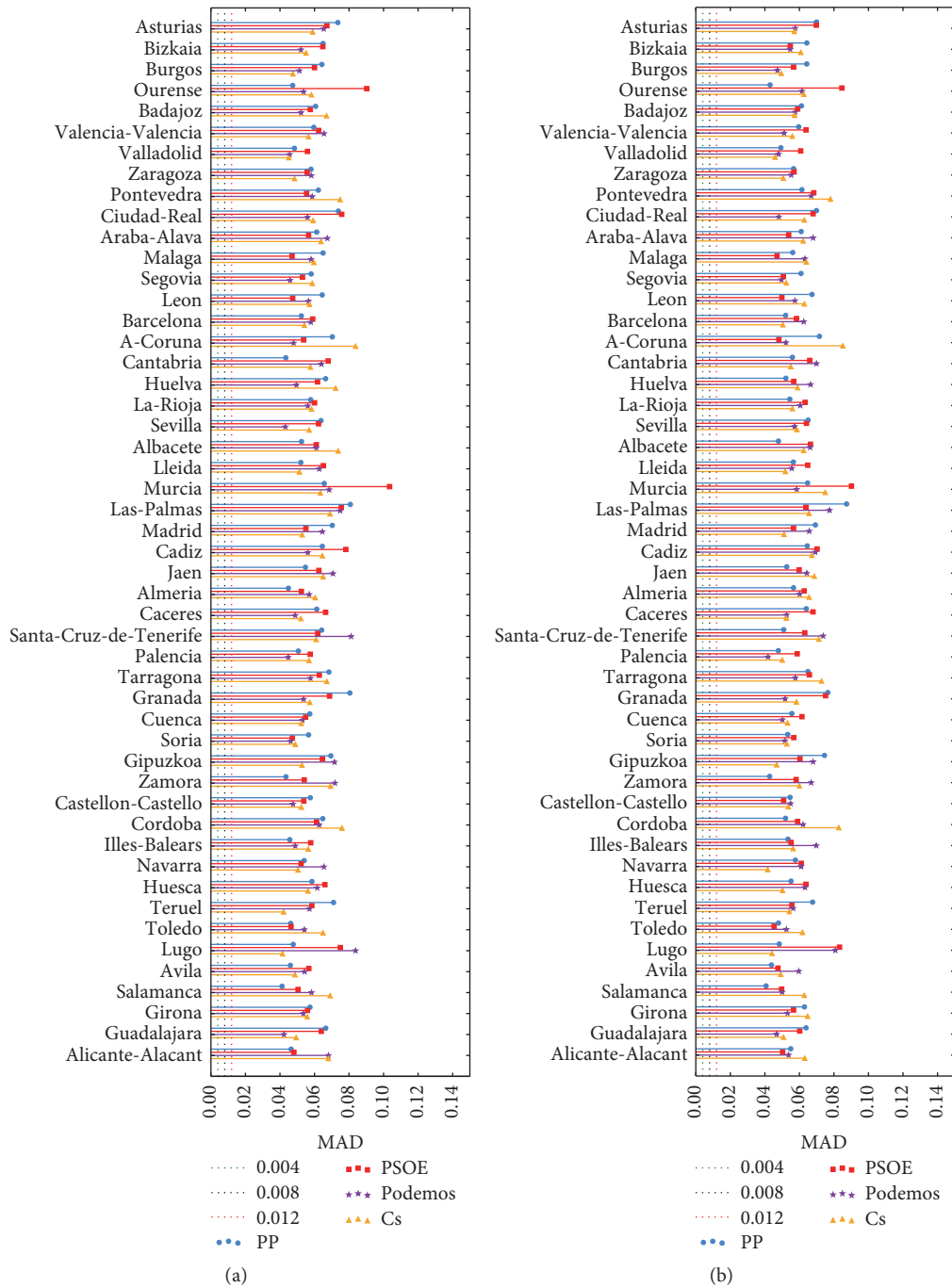


FIGURE 20: MAD values of the goodness of fit to 2BL for 2015 (a) and 2016 (b) at the aggregation level of precincts. In every case, the critical values for rejection at the 95 and 99% confidence level are shown. Virtually in all cases the null hypothesis is rejected.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors' Contributions

Juan Fernández-Gracia and Lucas Lacasa designed the study, Juan Fernández-Gracia performed the data gathering and

analysis, Juan Fernández-Gracia and Lucas Lacasa interpreted the results, Lucas Lacasa wrote the manuscript, and Juan Fernández-Gracia and Lucas Lacasa revised the manuscript.

Acknowledgments

The authors thank M. Antònia Tugores for her helpful assistance in the data gathering process, Luis F. Lafuerza

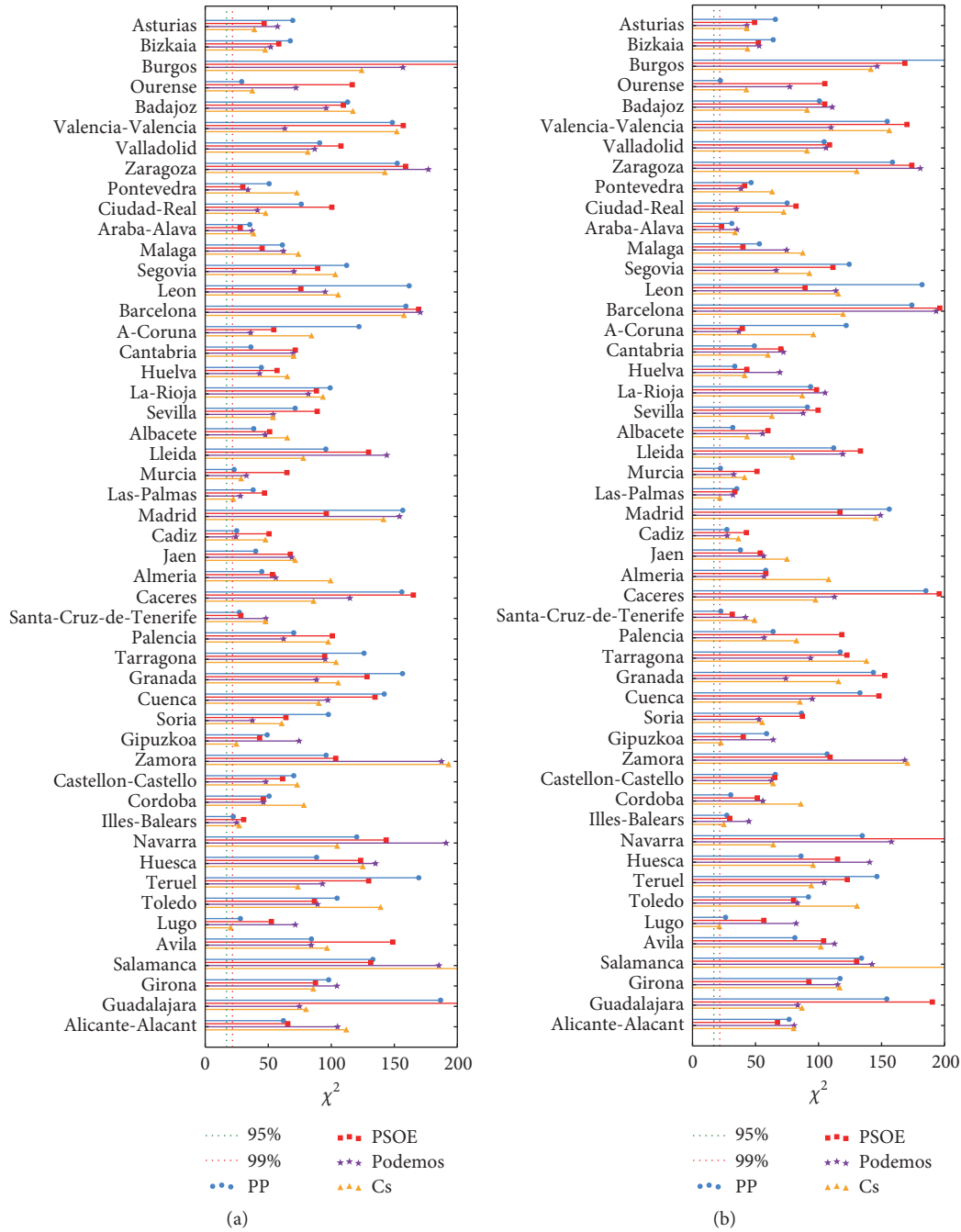


FIGURE 21: χ^2 values of the goodness of fit to 2BL for 2015 (a) and 2016 (b) at the aggregation level of precincts. In every case, the critical values for rejection at the 95 and 99% confidence level are shown. Virtually in all cases the null hypothesis is rejected. All these results are consistent with the hypothesis test based on MAD reported in Figure 20.

for fruitful suggestions, and D. Kobak for pointing out the recent works [17, 18]. Lucas Lacasa acknowledges funding from EPSRC Early Career Fellowship EP/P01660X/1.

References

[1] R. M. Alvarez, T. E. Hall, and S. D. Hyde, *Election Fraud: Detecting and Deterring Electoral Manipulation*, Brookings Institution Press, Washington, Wash, USA, 2008.

[2] <http://www.infoelectoral.interior.es/min/>.
 [3] M. J. Nigrini, *Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection*, John Wiley and Sons, New Jersey, NY, USA, 2012.
 [4] W. R. Mebane, *Election Forensics: Vote Counts and Benford's Law*, Political Method ology Society, University of California, California, Calif, usa, 2006.
 [5] P. Klimek, Y. Yegorov, R. Hanel, and S. Thurner, "Statistical detection of systematic election irregularities," *Proceedings of*

- the National Academy of Sciences of the United States of America*, vol. 109, no. 41, pp. 16469–16473, 2012.
- [6] M. O. Hill, “Diversity and evenness: a unifying notation and its consequences,” *Ecology*, vol. 54, no. 2, pp. 427–432, 1973.
- [7] V. Latora, V. Nicosia, and G. Russo, *Complex Networks: Principles, Methods, and Applications*, Cambridge University Press, Cambridge, UK, 2017.
- [8] M. Rosvall and C. T. Bergstrom, “Maps of random walks on complex networks reveal community structure,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 4, pp. 1118–1123, 2008.
- [9] D. Edler and M. Rosvall, “The MapEquation software package,” <http://www.mapequation.org>.
- [10] F. Benford, “The law of anomalous numbers,” *Proceedings of the American Philosophical Society*, vol. 78, pp. 551–572, 1938.
- [11] T. P. Hill, “A statistical derivation of the significant-digit law,” *Statistical Science*, vol. 10, no. 4, pp. 354–363, 1995.
- [12] B. Luque and L. Lacasa, “The first-digit frequencies of prime numbers and Riemann zeta zeros,” *Proceedings of the Royal Society A Mathematical, Physical and Engineering Sciences*, vol. 465, no. 2107, pp. 2197–2216, 2009.
- [13] R. Mansilla, “Análisis de los resultados electorales a partir de la ley de Benford,” <http://www.fisica.unam.mx/octavio>.
- [14] B. F. Roukema, *Benford’s Law Anomalies in the 2009 Iranian Election*, Torun Centre for Astronomy: Nicolaus Copernicus University, Torun, Poland, 2009.
- [15] T. P. Hill, “The significant-digit phenomenon,” *The American Mathematical Monthly*, vol. 102, no. 4, pp. 322–327, 1995.
- [16] C. Breunig and A. Goerres, “Searching for electoral irregularities in an established democracy: applying Benford’s Law tests to Bundestag elections in Unified Germany,” *Electoral Studies*, vol. 30, no. 3, pp. 534–545, 2011.
- [17] D. Kobak, S. Shpilkin, and M. S. Pshenichnikov, “Statistical fingerprints of electoral fraud?” *Significance*, vol. 13, no. 4, pp. 20–23, 2016.
- [18] D. Kobak, S. Shpilkin, and M. S. Pshenichnikov, “Integer percentages as electoral falsification fingerprints,” *The Annals of Applied Statistics*, vol. 10, no. 1, pp. 54–73, 2016.
- [19] J. Deckert, M. Myagkov, and P. C. Ordeshook, “Benford’s Law and the detection of election fraud,” *Political Analysis*, vol. 19, no. 3, pp. 245–268, 2011.
- [20] W. R. Mebane, “Comment on “benford’s law and the detection of election fraud”,” *Political Analysis*, vol. 19, no. 3, pp. 269–272, 2011.

Research Article

Spatial “Artistic” Networks: From Deconstructing Integer-Functions to Visual Arts

Ernesto Estrada ¹ and Puri Pereira-Ramos²

¹Department of Mathematics & Statistics, University of Strathclyde, 26 Richmond Street, Glasgow G1 1HQ, UK

²PeRArt Studio, Murgas 4, 15822 A Coruna, Spain

Correspondence should be addressed to Ernesto Estrada; ernesto.estrada@strath.ac.uk

Received 19 September 2017; Accepted 9 December 2017; Published 17 January 2018

Academic Editor: Gerard Olivar-Tost

Copyright © 2018 Ernesto Estrada and Puri Pereira-Ramos. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Deconstructivism is an aesthetically appealing architectonic style. Here, we identify some general characteristics of this style, such as decomposition of the whole into parts, superposition of layers, and conservation of the memory of the whole. Using these attributes, we propose a method to deconstruct functions based on integers. Using this integer-function deconstruction we generate spatial networks which display a few artistic attributes such as (i) biomorphic shapes, (ii) symmetry, and (iii) beauty. In building these networks, the deconstructed integer-functions are used as the coordinates of the nodes in a unit square, which are then joined according to a given connection radius like in random geometric graphs (RGGs). Some graph-theoretic invariants of these networks are calculated and compared with the classical RGGs. We then show how these networks inspire an artist to create artistic compositions using mixed techniques on canvas and on paper. Finally, we call for avoiding that the applicability of (network) sciences should not go in detriment of curiosity-driven, and aesthetic-driven, researches. We claim that the aesthetic of network research, and not only its applicability, would be an attractor for new minds to this field.

1. Introduction

There are multiple connections between networks and the visual arts. The study of graph drawing is an old topic in computer sciences and one of its main goals is the representation of networks in aesthetically appealing ways [1, 2]. In modern network theory, there have been extraordinary advances in the visualization of giant complex networks, which can be considered as pieces of art by themselves [3]. A different direction is the use of networks as an artistic mean of expression. The artistic work of Tomás Saraceno is an example of this kind of symbiosis where the author has used spider webs to create a universe of expressions [4]. Other artists melt networks into evocative images of the real-world to produce artistic designs. This is the case of the artist J. K. Rofling who has produced many of these symbiotic images [5]. Some examples of the work of J. K. Rofling are illustrated in Figure 1.

Here, we explore a different approach to connect networks and the visual arts. Essentially, we start from the construction of spatial networks based on simple rules, namely,

the location of points in a unit square. However, the coordinates of these points are generated by a mathematical transformation of integer numbers that generates artistic patterns on the plane. The inspiration for such transformation of integers and functions based on them comes from the “poststructuralist” school of philosophy and literary criticism known as deconstruction. This school started in the late 1960 after the influential book *De La Grammatologie* (1967) by the French philosopher Derrida [6]. This school of philosophical thinking influenced any areas of intellectual and creative activity including novels, poetry, architecture, the fine arts, and music. In architecture in particular, the term “deconstructivism” was adopted since the end of the 1980s [7]. According to Derrida this architectural style “*is not simply the technique of an architect who knows how to deconstruct what has been constructed but a probing which touches upon the technique itself, upon the authority of the architectural metaphor and thereby constitutes its own architectural rhetoric*” (cited by Hoteit in [7]).

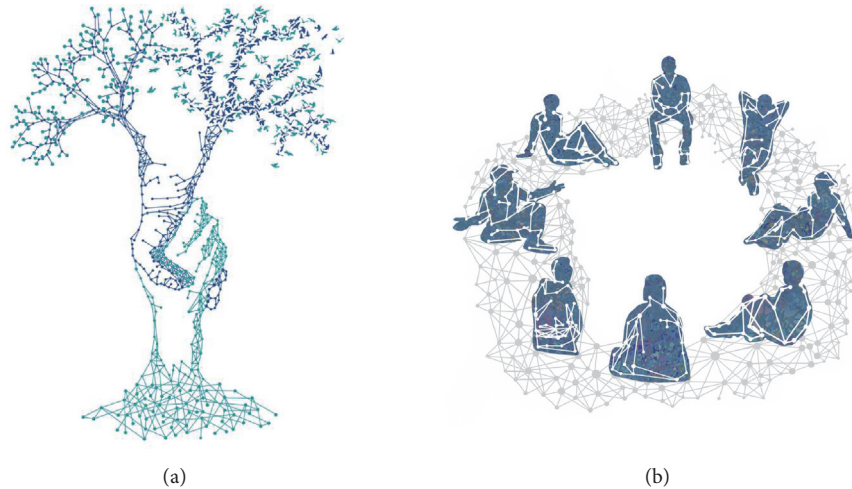


FIGURE 1: Two of the works produced by J. K. Rofling and taken from [5] with permission of the artist. (a) The Trand. (b) The Guys.

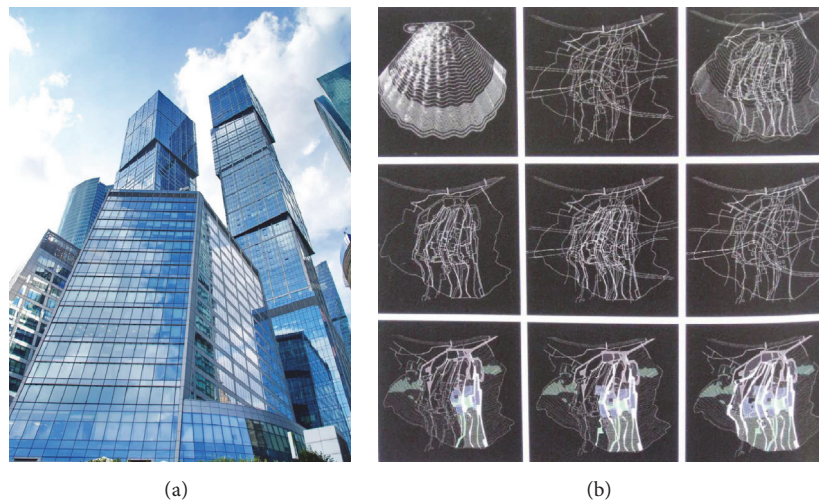


FIGURE 2: (a) City of Capitals in Moscow IBC, Russia. (b) Diagram of City of Culture of Galicia, Santiago de Compostela, Spain, by Peter Eisenman.

We do not pretend here to make a complete analysis of the deconstructivism in architecture but mainly of having a basic idea of its principles to be applied beyond its original frontiers. We then notice that the architectural deconstructivism looks initially as a fragmentation of the buildings which lack any visual logic. However, the deconstructing work accounts not only for this fragmentation but also for keeping a “memory” of the original composition in such a way that it “remembers” what it was in the beginning, that is, a building. In the City of Capitals in Moscow IBC, Russia, which is illustrated in Figure 2(a), the building is deconstructed into its unit block, that is, a cube, which is then “multiplied” to create again a tower with a different shape as the traditional ones. Another characteristic of deconstructivism is that the whole work must superimpose elements in such a way that “the design is produced, and the idea follows as its result.” As described by Hoteit [7] one of these examples is the City of Culture of Galicia, Santiago de Compostela, Spain, designed

by Eisenman. According to Hoteit [7] “Eisenman was mostly known for using the superimposition of layers.” In his creation of the City of Culture of Galicia “Eisenman determined the following four local traces: The downtown’s historical street grid; the typography of a hill; the abstract Cartesian grid; and the symbol of the city of Santiago, which is the scallop shell. Then, he superimposed these four abstracted traces to create an imaginary site condition, which became a real site for his project” (see also [8]). This idea is illustrated in Figure 2(b).

The connection with mathematical ideas here is evident. The superposition of layers can be imitated by the sum of parts and the compositional part can be obtained by multiplying the deconstructed parts such that we can recover certain “memory” from the original object. There are of course several ways of imitating these two characteristics of deconstructivism, but we have selected these two for the sake of mathematical convenience. This idea attempts to follow the existing line of connection between mathematical objects and

visual arts. This includes among others knots [9, 10], mosaics and tiles [11, 12], Fourier series [13], topological tori [14], and fractal curves [15], all of which produce artistic patterns of undoubtful beauty by themselves.

It can also be argued that some works in the cubism movement show elements of deconstruction. Indeed, analytic cubism is seen as an influential stream for deconstructivism via the work of Frank Gehry. Analytical cubism includes important paintings by Picasso, Braque, Metzinger, and others [16, 17]. Here, again, the principles of fragmenting, integrating, and superimposing are relevant in the analysis of these works [17]. Focusing only on these three principles to understand deconstruction is a clear oversimplification. However, we consider them here as the angular stone for what we will consider in the current work. Here, we are concerned with a formulation of deconstruction principles in mathematics.

2. Deconstructing Integer-Based Functions

Formulating deconstructivist principles for the whole of mathematics is a too ambitious project for a single paper. Instead, we focus here on integers and functions of integers. Then, the question is how to deconstruct an integer? The first idea should be to consider the individual digits of an integer as its building blocks. That is, for an integer x written in a given base b , it is represented by

$$x = a_1 b^n + a_2 b^{n-1} + a_3 b^{n-2} + \cdots + a_{n-1} b + a_n, \quad (1)$$

where $a_i \in \mathbb{Z}$ are nonnegative integers, which can be considered as the building blocks of x . For instance, the building blocks of $x = 2018$ are 2, 0, 1, and 8. Here the “whole” is represented by the integer, which in architecture should be the tower. The blocks are the digits forming that whole, like the cubes in the tower.

Now, we should proceed to the “superposition of layers” part. Here, we simply consider the function that sums the digits of the integer x in the base b [18]:

$$S_b(x) = \sum_{i=1}^n a_i = \sum_{k=0}^{\lfloor \log_b x \rfloor} \frac{1}{b^k} (x \bmod b^{k+1} - x \bmod b^k). \quad (2)$$

For instance, for $x = 2018$, the integration will produce $S_{10}(x) = 11$. These sequences for different bases b are stored in the *On-Line Encyclopaedia of Integer Sequences* [19, 20]; for instance, A007953 is the sequence for $b = 10$.

In order to complete the deconstruction of the integer we need the “recovery of the memory” of the original object. That is, we consider the product of the integer x by $S_{10}(x)$ as the final deconstruction of the integer x [21]:

$$\widehat{x}_b = x \sum_{i=1}^n a_i. \quad (3)$$

In this way, we have that a given integer is first dismembered into its digits; then the digits are superimposed to each other as the different layers of the integer using the digit-sum function. Finally, we “recover” the memory of the original number by multiplying the integer by its digit-sum. Hereafter,

we consider only the base $b = 10$; thus $\widehat{x} = \widehat{x}_b$. Using this approach, the deconstructed integers “remember” something about their original numbers. For instance, $\widehat{19} = 190$, $\widehat{28} = 280$, $\widehat{37} = 370$, $\widehat{46} = 460$, $\widehat{55} = 550$, $\widehat{91} = 910$, and $\widehat{82} = 820$ (see sequence A117570 in [19]). However, it does not mean that \widehat{x}_b is different for each integer. For instance, $\widehat{75} = \widehat{150} = 900$.

Let us now extend this approach to any function based on integers. Let $f(x)$ be a function of the number x , for example, $\sin(x)$. Then, the sum of digit-functions $\widehat{f}(x) : \mathbb{Z} \rightarrow \mathbb{R}$ as the function defined on the integers, such that

$$\widehat{f}(x) = (f(a_1) + f(a_2) + \cdots + f(a_n)) f(x). \quad (4)$$

For negative integers $-x$, if the function $f(-x)$ exists, we define

$$\widehat{f}(-x) = (f(a_1) + f(a_2) + \cdots + f(a_n)) f(-x). \quad (5)$$

We then consider the plot of pairs of functions $\widehat{f}(t)$ and $\widehat{g}(t)$ for the integers $t \leq n/2$ such that

$$\begin{aligned} x &= \widehat{f}(t), \\ y &= \widehat{g}(t). \end{aligned} \quad (6)$$

If the functions $\widehat{f}(t)$ and $\widehat{g}(t)$ are also defined for negative arguments we obtain the corresponding transforms for $-n/2 \leq t$. We are going to use these functions to build spatial networks as described in the next section.

3. Building Spatial Networks

In this section, we define our strategy for building spatial graphs based on the deconstruction of integer-functions. This strategy is based on the random geometric graphs (RGGs). Thus, we first explain the way in which RGGs are built. The RGG is defined by distributing uniformly and independently n points in the unit d -dimensional cube $[0, 1]^d$ [22]. Hereafter we consider only the 2-dimensional case. Then, two points are connected by an edge if their Euclidean distance is at most R , which is a given fixed number known as the *connection radius*. That is, we create a disk of radius R centered at each node, and every node inside that disk is connected to the central node as illustrated in Figure 3. A few important structural parameters of RGGs have been determined analytically in the literature (see, e.g., [22]).

Now, let us consider the process T that generates n points in the unit square according to the transforms of integer-functions defined in the previous section. For instance, let us consider $-1000 \leq t \leq 1000$ and make the following transformation.

Transform 1 (T_1).

$$\begin{aligned} x &= \widehat{t}, \\ y &= \begin{cases} \widehat{\sin t}, & t \leq 0 \\ -\widehat{\sin t}, & t > 0. \end{cases} \end{aligned} \quad (7)$$

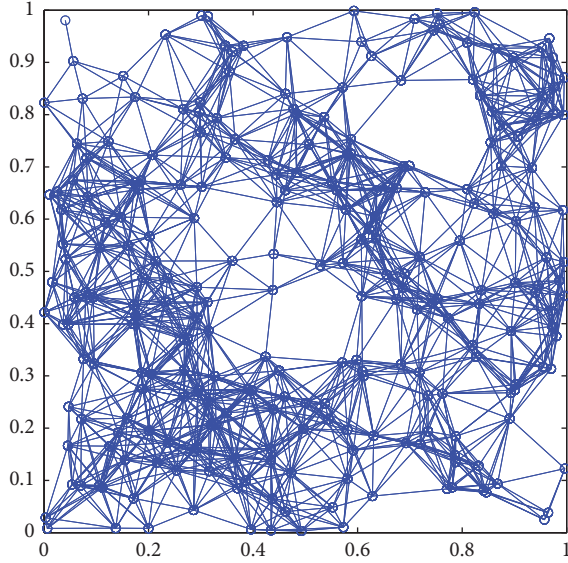


FIGURE 3: Illustration of a RGG created with 250 nodes embedded into a unit square where the nodes are connected if they are at a Euclidean distance smaller than or equal to $R = 0.15$.

Notice that we consider the trigonometric functions of the numbers in degrees not in radians. For instance, $\sin t$ means “sine of t degrees.” Then, we plot every point on the unit square according to its coordinates (x, y) defined before as illustrated in Figure 4(a). Using the approach to construct RGGs described before we construct the network for a given value of R . That is, after placing the points in the unit square we center a disk of radius R on each point and connect to it every other point which is inside the corresponding disk. Here we will use radii which guarantee the connectivity of the graph—the study of the connectivity of these graphs is beyond the scope of the current work. For instance, in Figure 4(b) we illustrate the network created by using $R = 0.075$.

4. Spatial “Artistic” Networks

It is straightforward to realize that the previously obtained spatial graph (Figure 4(b)) displays a few artistic attributes: (i) biomorphic shape, that is, suggestive in shape of a living organism (a butterfly in this case); (ii) symmetry; and (iii) beauty, just to mention three. The appearance of a biomorphic shape here is just by chance and we have selected in this work only those transforms of integer-functions which produce artistically appealing shapes. However, it must be emphasized that both—beauty and interpretation of shapes—are on the eyes of the beholder, and different observers can see different things in these and other spatial networks created from integer-functions. Here we coin the name *spatial “artistic” networks* (SANs) for the spatial networks created using the previously described method.

Let us now consider other alternatives to the integer-function deconstruction to see which artistic objects we can obtain. Artistic composition is the result of artist creativity and it includes a series of general rules that can be

implemented computationally. Here, we mainly follow a handmade compositional creation in order to glue series of integer-function transforms into single art works. For instance, let us consider the following parametric equations.

Transform 2 (T_2).

$$\begin{aligned} x &= \widehat{t}, \\ y &= \widehat{\cos t} + \widehat{t} \cdot \widehat{\sin t}. \end{aligned} \quad (8)$$

The resulting network for $R = 0.085$ with $n = 1000$ points is illustrated in Figure 5 where we have used $-500 \leq t \leq 500$ and the nodes are colored according to their closeness centrality.

Transform 3 (T_3). Another example is obtained by transforming the Astroid curve using the integer-digit transform. First, let us remind the reader that the Astroid is the curve: $x = \cos^3(t)$ and $y = \sin^3(t)$. Then, we make the transformation of the coordinates as explained before, such that we have

$$\begin{aligned} x &= (\widehat{\cos t})^3, \\ y &= (\widehat{\sin t})^3. \end{aligned} \quad (9)$$

The corresponding SAN is illustrated in Figure 6, where we have used again $-1000 \leq t \leq 1000$ and the nodes are colored according to their closeness centrality.

Transform 4 (T_4). The involute of the circle— $x = \cos t + t \sin t$; $y = \sin t - t \cos t$ —can also be transformed accordingly for $-1000 \leq t \leq 1000$ such that we obtain the following parametric equations:

$$\begin{aligned} x &= \widehat{\sin t} - \widehat{t} \widehat{\cos t}, \\ y &= -\widehat{\cos t} - \widehat{t} \widehat{\sin t}, \end{aligned} \quad (10)$$

which produce the network illustrated in Figure 7.

Transform 5 (T_5). Finally, we obtain the integer-function transformation of the cardioid curve, such that

$$\begin{aligned} x &= 2\widehat{\cos t} + \widehat{\cos(2t)}, \\ y &= \frac{1}{2}\widehat{\cos t} - 5\widehat{\sin(2t)}, \end{aligned} \quad (11)$$

where

$$\begin{aligned} \widehat{\cos(2t)} &= (\cos(2a_1) + \cos(2a_2) + \cdots + \cos(2a_n)) \cos(2t) \\ \widehat{\sin(2t)} &= (\sin(2a_1) + \sin(2a_2) + \cdots + \sin(2a_n)) \sin(2t), \end{aligned} \quad (12)$$

with $t = a_1 b^n + a_2 b^{n-1} + a_3 b^{n-2} + \cdots + a_{n-1} b + a_n$ represented in the decimal basis $b = 10$. The spatial graph based on this transformation is illustrated in Figure 8 where we have used $-1000 \leq t \leq 1000$.

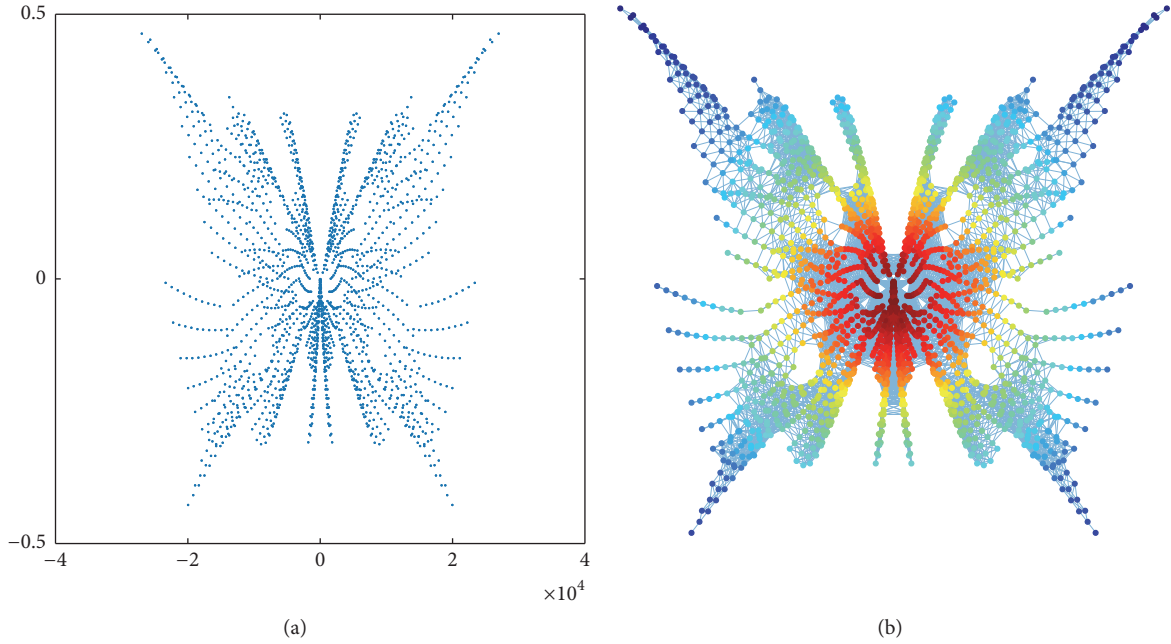


FIGURE 4: Illustration of the process to build a spatial network based on integer-function deconstruction. (a) Distribution of the points obtained from the transform T_1 on a square. (b) Construction of the spatial graph using a connection radius $R = 0.075$ with $n = 2,000$ points and coloring the nodes according to their closeness centrality.

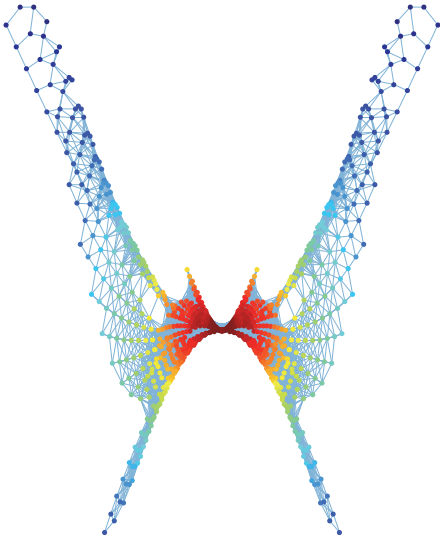


FIGURE 5: Spatial network constructed from the distribution of points in a unit square according to the transform T_2 using a connection radius $R = 0.085$ with $n = 1000$ points and coloring the nodes according to their closeness centrality.

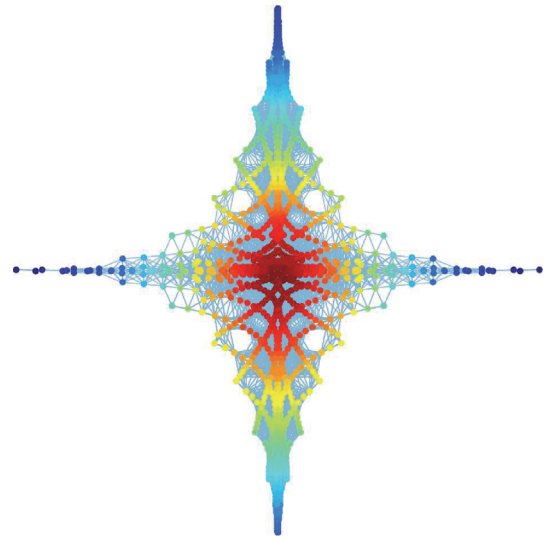


FIGURE 6: Spatial network constructed from the distribution of points in a unit square according to the transform T_3 using a connection radius $R = 0.1$ with $n = 2000$ points and coloring the nodes according to their closeness centrality.

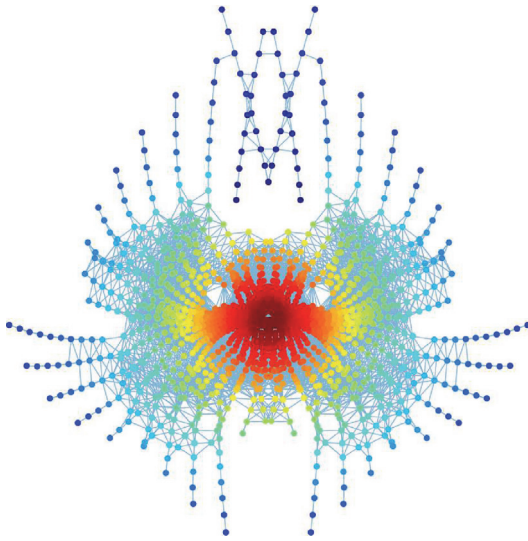
5. Network Invariants of SANs

Here we consider a few invariants of the networks constructed by using the five transformations previously studied and compare them with the same invariants for the analogous RGG. That is, we construct RGGs with the same number of nodes and connection radius than the SANs created by the previously defined transforms. These invariants are as

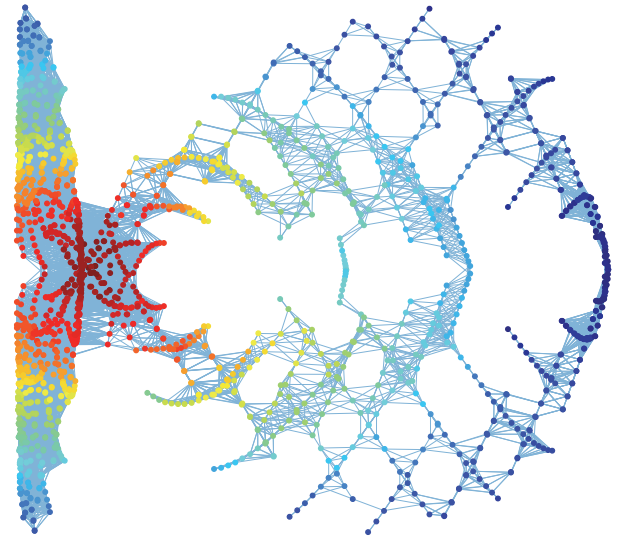
follows: the number of nodes n , the number of edges m , the edge density δ , the maximum degree k_{\max} , the average Watts-Strogatz clustering coefficient \bar{C} , the global transitivity index C , average shortest path distance \bar{d} , network diameter d_{\max} , and the degree assortativity r (for definitions and meaning see [23]). In Table 1, we give the values of these graph-theoretic invariants for the SANs and RGGs studied here.

TABLE 1: Graph-theoretic invariants of the spatial “artistic” networks described in Section 4.

	T_1		T_2		T_3		T_4		T_5	
	SAN	RGG	SAN	RGG	SAN	RGG	SAN	RGG	SAN	RGG
R	0.075		0.085		0.1		0.085		0.1	
n	2,000		1,000		2,000		2,000		2,000	
m	30,158	33,128	32,286	10,440	129,656	57,287	198,951	42,113	61,217	57,287
δ	0.015	0.017	0.077	0.021	0.065	0.029	0.099	0.021	0.031	0.029
k_{\max}	117	55	205	39	299	87	581	64	150	87
\bar{C}	0.618	0.613	0.665	0.608	0.727	0.622	0.642	0.614	0.695	0.622
C	0.217	0.200	0.249	0.198	0.247	0.202	0.264	0.202	0.225	0.202
\bar{d}	11.16	8.22	7.38	7.59	7.45	6.14	6.08	7.266	11.90	6.14
d_{\max}	46	21	39	19	23	15	34	18	33	15
r	0.65	0.60	0.75	0.59	0.74	0.61	0.79	0.60	0.68	0.61

FIGURE 7: Spatial network constructed from the distribution of points in a unit square according to the transform T_4 using a connection radius $R = 0.085$ with $n = 2000$ points and coloring the nodes according to their closeness centrality.

In general, the graph-theoretic properties of SANs are relatively similar to those of the RGGs. However, there are some differences, particularly for the maximum degree and maximum distance. That is, the SANs always have significantly larger k_{\max} and d_{\max} than the corresponding RGGs. These two parameters are larger in the SANs as a consequence of the higher concentration of points in the center of the figure in relation to their peripheries. This situation is avoided in the RGG due to the random and homogeneous distributions of the points in the unit square. The similarities in terms of clustering coefficients and assortativity—notice that all networks are degree assortative—between SANs and RGGs are remarkable. We, however, are not claiming any application of these graphs for solving problems in the real-world, apart from being a source of artistic inspiration. Then, the analysis of these properties is mostly a curiosity-driven one and not the search for useful properties of these graphs. In the next section, we explore how these networks inspire some art.

FIGURE 8: Spatial network constructed from the distribution of points in a unit square according to the transform T_5 using a connection radius $R = 0.1$ with $n = 2,000$ points and coloring the nodes according to their closeness centrality.

6. Artistic Inspiration

Science is sometimes seen as a dry and cold activity, such that it is not able to inspire those which are not involved in it. In earlier definitions of the humanities as “*the branches of polite learning, especially the ancient classics and literature of aesthetics, as distinguishes from informational or utilitarian values*” the sciences are marginalized as “*informational but unaesthetic, that is, as useful but grubby*” [24]. Many efforts are currently done for attracting the attention of the general public to the beauty of scientific discoveries. In mathematical sciences, for instance, there are initiatives, such as Bridges [25, 26], which bring together mathematicians and artists to produce artistic works from, or inspired by, mathematics. The *Journal of Humanistic Mathematics* [27] has also been launched to fill the gap between the humanities and mathematics.

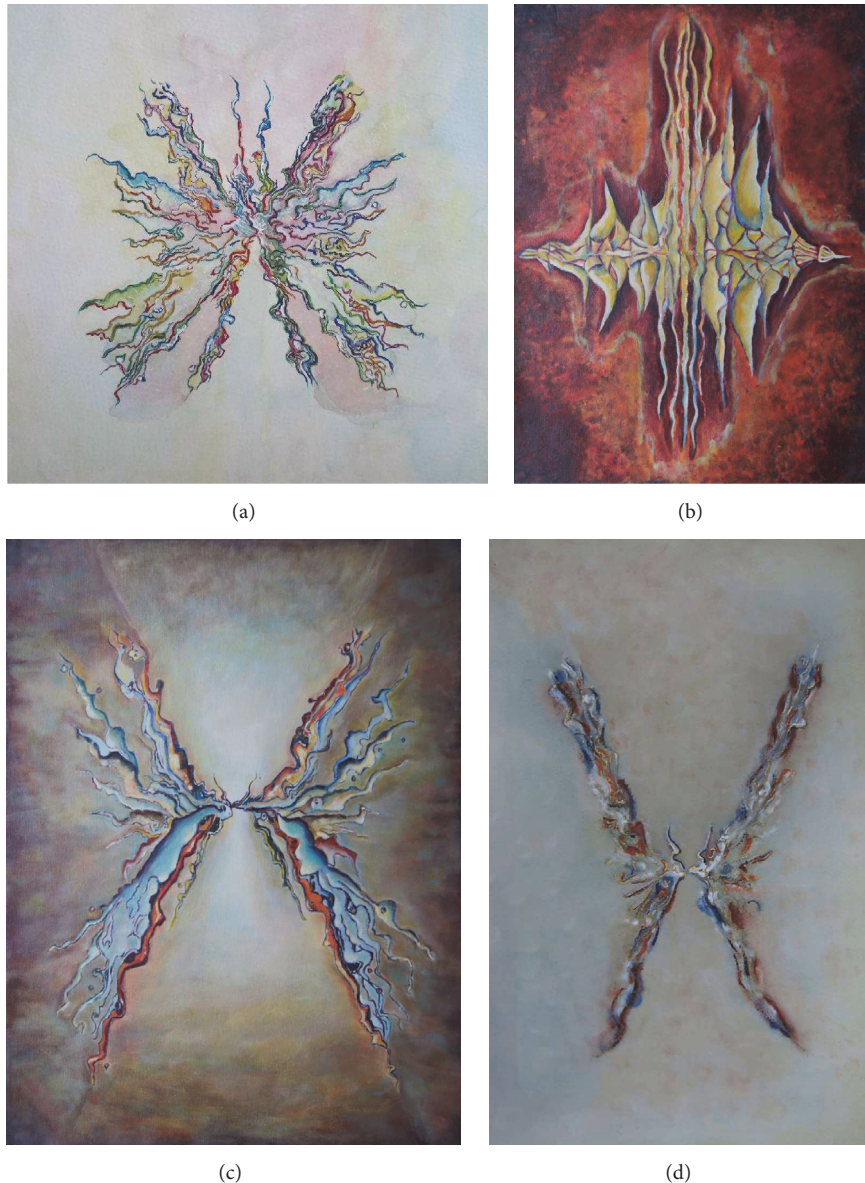


FIGURE 9: Photograph of four artistic works of artist Puri Pereira. (a) “Butterfly # 1” painted in acrylic, ink, and watercolor on paper of dimensions 30×40.5 cm. (b) “Fire on Water” painted in acrylic on canvas of dimensions 45.7×61 cm. (c) “Butterfly # 2” painted in acrylic on canvas of dimensions 45.7×61 cm. (d) “Birds” painted in ink, watercolor, and acrylic on paper of dimensions 30×40.5 cm.

In this part of our work, we present a few snapshots of what an artist can bring from the visual images produced by the SANs obtained from the integer-functions deconstruction presented here. That is, the SANs previously described have been the source of artistic inspiration for the production of purely aesthetic works outside the constraints of (network) sciences. The results are illustrated in Figure 9.

7. On the Artistic Value of (Network) Sciences

Network sciences have an important impact on our understanding of nature and modern society. Its practical importance has been documented in many papers in the last few years. But network science is also driven by aesthetic criteria.

Sometimes it is the mathematical beauty of the equations describing the structure of, or the dynamics on, the networks. Sometimes it is the beauty of the embedding of the network into certain space that produces outstanding visualizations. Other times it is the result of the application of network theory to a particular problem that produces an aesthetic feeling due to the beauty of the findings or what is unexpected of the connections found. Then, the importance of the applications of networks to solve practical problems should not hide its inherent beauty. The applicability of (network) sciences should not go in detriment of curiosity-driven, and aesthetic-driven, researches. We should find a compromise between application-driven and curiosity-driven researches. Abraham Flexner [28]—who was a founder of the Institute of

Advanced Studies in Princeton and its Director from 1930 to 1939—stressed that “institutions of learning should be devoted to the cultivation of curiosity and the less they are deflected by considerations of immediacy of application, the more likely they are to contribute not only to human welfare but to the equally important satisfaction of intellectual interest which may indeed be said to have become the ruling passion of intellectual life in modern times.” Obviously, there are many pressing problems in modern society that we are aimed to solve using network methods and approaches, and we should never forget our social responsibility. But our institutions should not forget either that as Flexner remarked “a poem, a symphony, a painting, a mathematical truth, a new scientific fact, all bear in themselves all the justification that universities, colleges, and institutes of research need or require” [28]. Thus, we should be reminded that (network) science has a humanistic side, which is as important as the many applications that it has found. Forgetting this side of it—its beauty and capacity of surprising—is similar to tear a wing to a bird. We all know that birds with only one wing cannot fly.

8. Conclusions

The spatial artistic networks (SANs) created here are the product of a curiosity-driven process more than of any practical necessity or real-world application. Thus, the value of these networks does not reside in their usefulness as a mathematical tool for modeling reality but as a source of inspiration of artistic work as well as attractive objects per se. We do not discard, however, that such networks can find some applications for modeling spatial processes in the real-world, due to their similarities with RGGs as well as by the fact that the points here are not randomly distributed in space but by using well-defined mathematical rules. The type of high-density core and very sparse periphery reminds one with the situation frequently found in many spatial networks like cities.

Many chemistry students have been motivated to their subject by the beauty of the representations of the molecular structure. In physics, a similar situation exists when we consider the aesthetic of cosmic landscapes and the structure of the universe. Mathematicians always claim to be seduced by the beauty of mathematical equations. Can we attract students into network sciences by appealing to the aesthetic beauty of networks? The only way to know it is by trying. We hope that the current work contributes to this goal, either by attracting curious minds to the field or by inspiring other researchers in the field to explore the beauty of networks per se.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

References

- [1] H. C. Purchase, R. F. Cohen, and M. James, “Validating graph drawing aesthetics,” in *International Symposium on Graph Drawing*, vol. 1027 of *Lecture Notes in Computer Science*, pp. 435–446, Springer, Berlin, Heidelberg, 1995.
- [2] H. C. Purchase, “Metrics for graph drawing aesthetics,” *Journal of Visual Languages and Computing*, vol. 13, no. 5, pp. 501–516, 2002.
- [3] M. De Domenico, M. A. Porter, and A. Arenas, “MuxViz: A tool for multilayer analysis and visualization of networks,” *Journal of Complex Networks*, vol. 3, no. 2, pp. 159–176, 2015.
- [4] <http://tomassaraceno.com/>.
- [5] <https://www.jkrofling.com/>.
- [6] J. Derrida, “De la Grammatologie,” *de Minuit*, 1967.
- [7] A. Hoteit, “Deconstructivism: Translation From Philosophy to Architecture,” *Canadian Social Science*, vol. 11, pp. 117–129, 2015.
- [8] V. Belogolovsky, “One-on-one: Architecture that leads to a point: Interview with Daniel Libeskind,” pp. 10–12, 2016, <http://www.archnewsnow.com/features/Feature369.htm>.
- [9] A. Åström and C. Åström, “Circular knotworks consisting of pattern no. 295: a mathematical approach,” *Journal of Mathematics and the Arts*, vol. 5, no. 4, pp. 185–197, 2011.
- [10] R. Bosch, “Simple-closed-curve sculptures of knots and links,” *Journal of Mathematics and the Arts*, vol. 4, no. 2, pp. 57–71, 2010.
- [11] R. Bosch and U. Colley, “Figurative mosaics from flexible Truchet tiles,” *Journal of Mathematics and the Arts*, vol. 7, no. 3–4, pp. 122–135, 2013.
- [12] X. Zheng and N. S. Brown, “Symmetric designs on hexagonal tiles of a hexagonal lattice,” *Journal of Mathematics and the Arts*, vol. 6, no. 1, pp. 19–28, 2012.
- [13] F. A. Farris, “Symmetric yet organic: Fourier series as an artist’s tool,” *Journal of Mathematics and the Arts*, vol. 7, no. 2, pp. 64–82, 2013.
- [14] C. H. Séquin, “Topological tori as abstract art,” *Journal of Mathematics and the Arts*, vol. 6, no. 4, pp. 191–209, 2012.
- [15] J. Briggs, *The patterns of chaos: A new aesthetic of art, science, and nature*, A Touchstone Book, Simon & Schuster, USA, 1992.
- [16] R. L. Taylor, “Cubism— Abstract or Realist?,” *Philosophy and the Visual Arts*, in *Cubism— Abstract or Realist*, pp. 77–95, Springer, Netherlands, Amsterdam, 1987.
- [17] D. A. Gall, “Fragments of what? Postmodernism, Hybridity and Collage,” *Journal of Art for Life*, vol. 5, p. 24, 2014.
- [18] L. E. Bush, “An asymptotic formula for the average sum of the digits of integers,” *The American Mathematical Monthly*, vol. 47, pp. 154–156, 1940.
- [19] N. J. Sloane and S. Plouffe, *The Encyclopedia of Integer Sequences*, Academic Press, 1995, <http://oeis.org>.
- [20] N. J. Sloane, “My favorite integer sequences,” in *Sequences and their applications*, C. Ding, T. Helleseth, and H. Niederreiter, Eds., pp. 103–130, Springer, 1998.
- [21] E. Estrada and L. A. Pogliani, “A new integer sequence based on the sum of digits of integers,” *Kragujevac Journal of Sciences*, vol. 30, pp. 45–50, 2008.
- [22] M. D. Penrose, *Random Geometric Graphs*, Oxford University Press, 2003.
- [23] E. Estrada, *The Structure of Complex Networks: Theory and applications*, Oxford University Press, 2012.
- [24] F. J. Rutherford, “A Humanistic Approach to Science Teaching,” *NASSP Bulletin*, vol. 56, no. 361, pp. 53–62, 1972.
- [25] K. Fenyvesi, “Bridges: A World Community for Mathematical Art,” *The Mathematical Intelligencer*, vol. 38, no. 2, pp. 35–45, 2016.
- [26] <http://www.bridgesmathart.org/>.
- [27] <http://scholarship.claremont.edu/jhm/>.
- [28] A. Flexner, *The Usefulness of Useless Knowledge*, Harper’s Magazine, 1939.

Research Article

Modeling and Simulation of Project Management through the PMBOK® Standard Using Complex Networks

Luz Stella Cardona-Meza¹ and Gerard Olivar-Tost²

¹Department of Information and Computation, Universidad Nacional de Colombia, Sede Manizales, Campus La Nubia, Manizales 170003, Colombia

²Department of Mathematics, Universidad Nacional de Colombia, Sede Manizales, Campus La Nubia, Manizales 170003, Colombia

Correspondence should be addressed to Gerard Olivar-Tost; golivart@unal.edu.co

Received 27 August 2017; Accepted 12 November 2017; Published 4 December 2017

Academic Editor: Dimitri Volchenkov

Copyright © 2017 Luz Stella Cardona-Meza and Gerard Olivar-Tost. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Discussion about project management, in both the academic literature and industry, is predominantly based on theories of control, many of which have been developed since the 1950s. However, issues arise when these ideas are applied unilaterally to all types of projects and in all contexts. In complex environments, management problems arise from assuming that results, predicted at the start of a project, can be sufficiently described and delivered as planned. Thus, once a project reaches a critical size, a calendar, and a certain level of ambiguity and interconnection, the analysis centered on control does not function adequately. Projects that involve complex situations can be described as adaptive complex systems, consistent in multiple interdependent dynamic components, multiple feedback processes, nonlinear relations, and management of hard data (process dynamics) and soft data (executive team dynamics). In this study, through a complex network, the dynamic structure of a project and its trajectories are simulated using inference processes. Finally, some numerical simulations are described, leading to a decision making tool that identifies critical processes, thereby obtaining better performance outcomes of projects.

1. Introduction

Projects have long been considered business practices of high value for organizations, with important results in general. Therefore, project management is considered a key factor for the success of projects and strategic objectives of companies [1]. In 1950, the social construct of *project management* was first introduced (in the United States Air Force). Its first proponent was Brigadier Bernard Schriever, who implemented the concept of concurrence, integrating all the elements of a project into a single program and budget, executed in parallel and not in sequence. Since then, specific techniques have arisen—histograms, chronograms, concepts of the life cycle of a project, and the work breakdown structure, which make up the knowledge base of the classic perspective of projects [2].

According to Padalkar and Gopinath [3], a significant part of the first studies on project management, which continued up to the 1980s, used conceptual or analytical methods.

These methods focused on the optimization of scheduling based on the premise that project activities and their interrelations were fixed and measurable [4–8].

Subsequently, in the 1990s, new results of empirical experiments studying the success and failure of projects surfaced [9–22]. Defining the concepts of success and failure associated with projects and their management is not an easy task, and there is no consensus on their definition or measurement. According to Baccharini [10], there must be a distinction between the success of a project as measured by the fulfillment of the requirements of the end product, and the success of the project management as measured habitually in terms of time, costs, and quality [23].

The pursuit of the success/failure of projects has led to an expansion of research on organization contexts that are broader, behavioral, and interdisciplinary [9, 14, 17, 24–28]. This has promoted other research on such topics as contingencies, behavior, and governance in projects, interrelations

between projects, decision making, and the perspective of complexity [2].

From the 1960s, a small deviation from the classic deterministic perspective started to emerge, in which nondeterministic processes began to be considered. This included criticism of PERT and the beta-distribution [29–31], and the treatment of project management as being not only deterministic [32]. The modeling of the uncertainty of project phenomena began to be considered as assumptions about attributes considered static broadened [33–39]. System dynamics began to be used for modeling the nonlinear effects of feedback loops in projects [40–43] and the modeling of projects under diffuse or probabilistic assumptions [44–48].

The dominant research focus has remained instrumentalist, with attempts to design models or methods of decision making with the goal of analyzing the performance of a project (e.g., the Project Management Body of Knowledge (PMBOK) standard from the Project Management Institute, which was founded in 1969 and its 5th edition was published in 2013).

The nondeterministic school of thought finds meaning in the weak theoretical nature of project management [1, 49–55].

From a brief exploration of literature, several attempts to model projects through assumptions related to complexity are found [56–61]. Several studies have theoretically discussed, defined, or provided constitutive elements of complexity [42, 50, 62–67, 67–73]. Other studies have done the same with regard to uncertainty [18, 60, 74–77].

In addition, the literature review provides evidence of the existence of several international organizations that have been expanding the body of knowledge of project management: ISO (21500), International Project Management Association, 1972, standard ICB 3.0, Association for Project Management, standard PRINCE2, Project Management Institute (PMI), 1969, PMBOK standard, International Centre for Complex Project Management, 2011, and New England Complex Systems Institute, amongst others.

For the purpose of this study and the evaluation of project management, the analysis is based on the PMBOK standard from the PMI, given its importance and international prevalence. It was chosen with the goal of describing an analytic structure of processes and as a tool for simulating complex networks used to evaluate project management nondeterministically.

1.1. Project Management as a Complex System. Complexity theory as applied to organizations [78] can also be applied to projects [42, 62]. All projects have attributes of interconnection, hierarchy, communication, control, and emergency, which are generally useful attributes for describing all types of systems [79]. In addition, most big and small projects exhibit characteristics of complex adaptive systems. They exhibit such characteristics as phase transitions, adaptability, and sensibility to initial conditions [79].

A complex project is a complex system made up of different elements interconnected to achieve an objective. Such a system can be described by a dynamic system, whose parts interact with each other and with their environment,

and such interactions give rise to new properties that did not previously exist [80].

According to [79], the most important characteristics exhibited by complex projects, seen as complex adaptive systems, are as follows:

- (i) Auto-organization: a project can suffer two types of perturbations—those of an exogenic nature (relating to changes of its environment) and those of an endogenous nature (relating to internal attribute changes that modify the relationships within the system) [81]. After a given perturbation, the project is reorganized until a new emerging structure is adopted, which can be fixed as long as no new environment or internal parameter changes occur.
- (ii) Hierarchy: projects as systems might contain other systems—the members of the temporary executive organization of the project in turn belong to other subsystems. In addition, the structures of work breakdown form hierarchies for the execution of activities, and the project can be perceived from different levels depending on the interest of the observer and so on.
- (iii) Nonlinearity: small perturbations cause effects in projects. The result of a small variation in the exogenous inputs or endogenous parameters can lead to considerable variations in the system (either immediately or in the future), contrasting a linear effect.
- (iv) Adaptability: adaptive systems can reorder their internal structure without the intervention of an external agent. This property, which is the product of unconscious learning, increases the probability that the system survives turbulent and unstable environments.

The remainder of this article is structured in the following manner. In Section 2, the modeling of process dynamics is described based on the PMBOK standard, along with the creation of a complex network. Section 3 is dedicated to numerical simulation and the results obtained. Finally, Section 4 describes the conclusions.

2. Modeling of a Complex Network of Processes Based on the PMBOK Standard

This section describes a possible algorithm to model complex project management through the creation of a complex network, in which nodes are the different processes of the project, and edges are the exchanges of information between such.

2.1. Determination of Generalities. Research, such as [82], suggests an appropriate sequence to develop a project management plan based on the PMBOK and focused on network theory. This research analyzes the activities of a project from a classic and deterministic perspective, while the present work evaluates the behavior of project management from the perspective of complexity, describing a structure for analyzing processes as a model that evaluates the dynamics of connections through simulation in a complex network, and

an analysis of the behavior of the characteristics of a complex project.

The methodological guide PMBOK, in its 5th edition (2013), describes five process groups and 10 knowledge areas that can be used to identify the relevant factors in arbitrary projects. The 10 knowledge areas are integration, scope, time, costs, quality, human resources, communications, risks, procurement, and stakeholders.

The five process groups are initiating, planning, executing, monitoring and controlling, and closing:

- (i) *Initiating process group* includes processes that define a new project or phase of an existing one, helping establish the vision and requirements. In this group, there are two subprocesses.
- (ii) *Planning process group* includes processes carried out to establish the scope of the plan, define and review the objectives, and develop an action plan to reach such objectives. In this group, there are 24 subprocesses.
- (iii) *Executing process group* includes processes targeted at completing the work defined in planning, with the goal of satisfying the requirements of such. In this group, there are eight subprocesses.
- (iv) *Monitoring and controlling process group* includes processes required to trace, analyze, and direct the progress and performance of the project, making necessary changes to it. In this group, there are 11 subprocesses.
- (v) *Closing process group* includes processes required to close a phase of the project or its entirety. In this group, there are two subprocesses.

2.2. Definition of the Analytic Structure of Processes. To model and simulate the analytic structure of processes, the following steps are defined.

- (i) Describe the subprocesses in terms of the flow of information.
- (ii) Define the model of connections between subprocesses.
- (iii) Determine parameters of the simulation.
- (iv) Present and discuss the results of the simulation of a complex network of subprocesses.

3. Simulation of the Complex Network of Processes or Subprocesses

In this section, subprocesses and the connections between them are described in terms of the flow of information. In addition, the parameters of the simulation are determined and the results of the simulations displayed.

3.1. Description of Subprocesses in terms of Information Flow.

(a) There are 26 initiating and planning subprocesses. They are the following: develop the project charter (see the example

in Figure 1), identify stakeholders, develop project management plan, plan scope management, collect requirements, define scope, create the EDT/WBS, plan schedule management, define activities, sequence the activities, estimate activity resources, estimate activity durations, develop schedule, plan cost management, estimate costs, determine budget, plan quality management, plan human resources management, plan communication management, plan risk management, identify risks, perform qualitative analysis, perform quantitative analysis, plan risk response, plan procurement management, and plan stakeholder management.

(b) There are 8 subprocesses in the executing process: direct and manage the work of the project (see an example in Figure 2), perform quality assurance, acquire project team, develop project team, manage project team, manage communications, conduct procurements, and manage stakeholder engagement.

(c) There are 11 subprocesses in the process of monitoring and controlling: monitoring and controlling project work, performing integrated change control, validating scope, controlling scope, controlling the schedule, controlling costs, controlling quality, controlling communications, controlling risks, controlling procurements, and controlling stakeholder engagement.

(d) There are two subprocesses in the closing process: close the project or phase and close the project or phase (procurements).

(e) Three new subprocesses are added (extending the PMBOK standard), and they are repository (contains the information of the project), exogenous (effects of the known environment variables), and novelties (effects of unknown variables).

As a result, 49 nodes are identified: 25 belonging to the subprocess of planning, 8 to the subprocess of execution, 11 to the subprocess of monitoring and controlling, 2 to the subprocess of closing, and 3 additional ones. In order to execute the activities of each subprocess, it is necessary for information to flow from other subprocesses and for new information to be generated, which flows into other subprocesses.

3.2. Connection Model between Subprocesses. Based on the subprocesses defined previously, relationships between these processes are defined as connections in the network, which are temporal. In addition, each subprocess is a node in the complex network.

The connections, or links, are established by the optimal relationship principle, which finds efficiencies between nodes and eliminates redundancies (i.e., not revisiting a node if it has been updated already). In order to elucidate this concept, consider the node “develop the project charter.”

- (i) The nodes that we label as sources are those that can send the following information: (1) exogenous node and/or (2) qualitative analysis of risk node.
- (ii) The nodes that we label as outputs are those that can receive information from (1) repository node, (2) developing project management plan, (3) planning scope management, (4) collecting requirements,

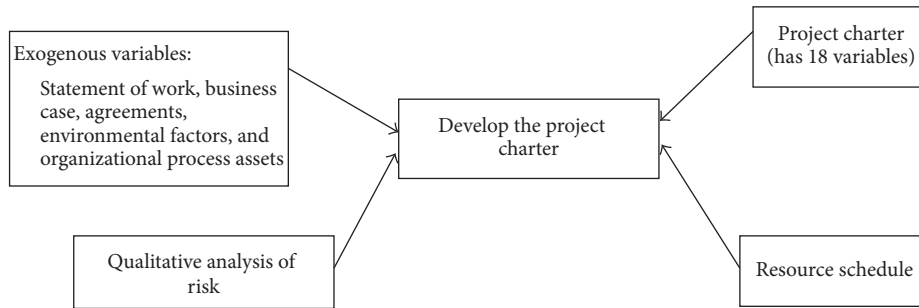


FIGURE 1: Area of knowledge of integration, initiation process, and development of the project charter subprocess (based on PMBOK).

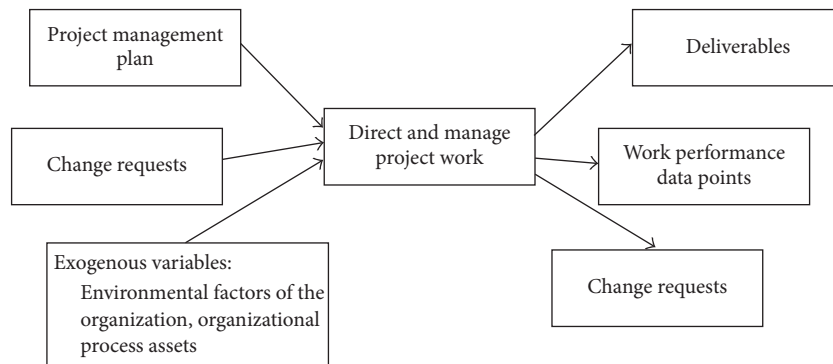


FIGURE 2: Area of knowledge of integration, executing process, and directing and management of project work subprocess (based on PMBOK).

(5) defining scope, (6) planning schedule management, (7) planning cost management, (8) planning risk management, and/or (9) identifying stakeholders.

When the node “develop the project charter” is activated by a source node, it processes the input information and then activates the output processes. These output processes in turn can activate other processes with which they have some relationship within the complex network. Thus, not only output node connections but also the secondary node sequence of the outputs is modeled.

3.3. Determining the Parameters of the Simulation. The conditions and hypothesis established in this subsection are the following, corresponding to the PMBOK standard.

- (i) The 49 subprocesses are determined based on PMBOK.
- (ii) It is assumed that there exists perfect information of the project in the planning phase and that the phases of executing, monitoring and controlling, and closing are carried out.
- (iii) The project considered is arbitrary, and the simulation attempts to evaluate the behavior of the project management are as defined by PMBOK.
- (iv) Each node is identified by a number, in order to simplify the simulation process.
- (v) Through color labeling, the dynamics of each node and connection can be established (see [83]):

- (a) green node, idle state
- (b) blue node, sending information state
- (c) yellow node, receiving information state
- (d) red node, processing information state
- (e) purple connection, active state sending information
- (f) black connection, inactive state

- (vi) The processing time of each node and the delay of the connection between two nodes are determined by a discrete uniform distribution that takes integer values between 1 and 11. This is done to exemplify the process, without setting the values to specific real cases, and these values are in the time units of the project.
- (vii) Connections are simulated concurrently. This takes place whenever a source node transfers information to several sinks, and when the sink node has several outputs (see [84–86]).
- (viii) The phases of project management are simulated in the following sequence: planning, executing, monitoring and controlling, and closing. This is also based on the PMBOK standard.

To illustrate the above, the subprocess “develop the project charter” is taken as an example, using a numerical identifier to simplify the simulation process, as follows: source nodes: exogenous (1) and qualitative analysis of risk

TABLE 1: Defining the connections of the network.

Source	Sink	Output
1; 38;	3;	0; 4; 9; 10; 11; 15; 22; 36; 46;

TABLE 2: Next generation of information connections between the nodes in the complex network.

Source	Sink	Output
4	9	0;
4	15	0;
...		
4	45	0;
4	45	1;
...		

(38); sink nodes: project charter (3); output nodes: repository (0), develop project management plan (4), plan scope management (9), collect requirements (10), define scope (11), plan schedule management (15), plan cost management (22), plan risk management (36), and identify stakeholders (46).

The aforementioned is described by Table 1, which contains 47 rows although only 1 is shown as an example.

Each output node can connect with other nodes and so on until the update/change of information is completed.

For example, when node 4, an output node, is updated/modified, it connects with the sink nodes, which in turn connect with the output nodes, and this process is repeated until all the related information is updated/modified (see Table 2). The table contains 403 rows although only a few are shown as an example.

The graphical representation of this dynamic, both of the nodes (which do not change) and the connections (which change with time), displays the different states as the complex network evolves. Using the previous example, this dynamic is depicted through the complex network represented in Figures 3, 4, 5, and 6, which form a sequence through time.

As depicted in this subsection, a change of states is observed, given by the dynamics of nodes and the dynamics of edges. For this example, activation starts at nodes 1 and 38, which process information; once they receive (red color), they then establish connections with node 3; once they do, nodes 1 and 38 send information (blue color), and node 3 receives information (yellow color), with an active connection (purple color). Then, node 3 processes the received information (red color). This process occurs repeatedly until connections ready for information processing are established.

Another way to understand the dynamics of nodes and connections, related to the previous example, is depicted in Figure 7 as a temporal network [87].

- (i) Initial state: nodes 1 and 38 process information concurrently (red color), in an amount of time determined by the discrete uniform distribution.
- (ii) Next state: nodes 1 and 38 transmit information (blue color), node 3 receives information (yellow color), and the connection between nodes 1-3 and 38-3 is active (purple color).

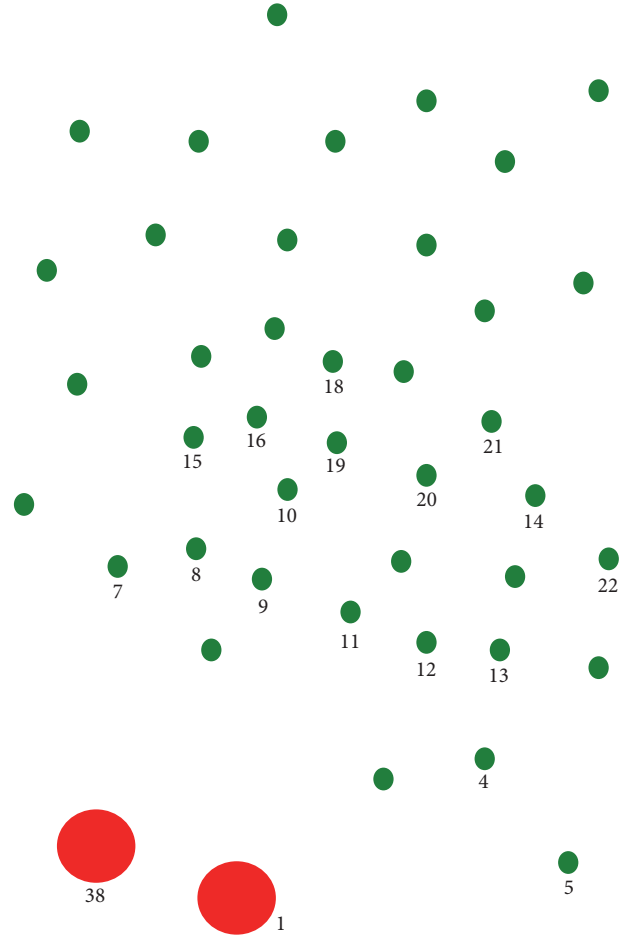


FIGURE 3: Nodes 1 and 38 process information.

- (iii) Next state: node 3 processes information (red color), nodes 1 and 38 are idle (green color), and connections between nodes 1-3 and 38-3 are inactive (black color).
- (iv) The process continues.

3.4. Results and Discussion of the Complex Network of Sub-processes Simulation. The results are described based on the simulations performed with the software created for that end. Not all the graphs can be displayed owing to space constraints.

The results of the simulation are as follows.

- (i) Based on the discrete uniform distribution considered previously for calculating times, a duration of 21,068 units of time for the planning phase and of 4,115 time units for the executing phase is obtained (summing to a total of 25,183 time units up to the end of this phase). The monitoring and controlling phase has a duration of 5,028 time units (summing to 30,211 time units); and the closing phase has a duration of 90 time units (giving a total of 30,301 time units). Thus, the total time taken by the simulation is 30,301 time units. As mentioned previously, a discrete uniform distribution taking values between 1 and 11 is

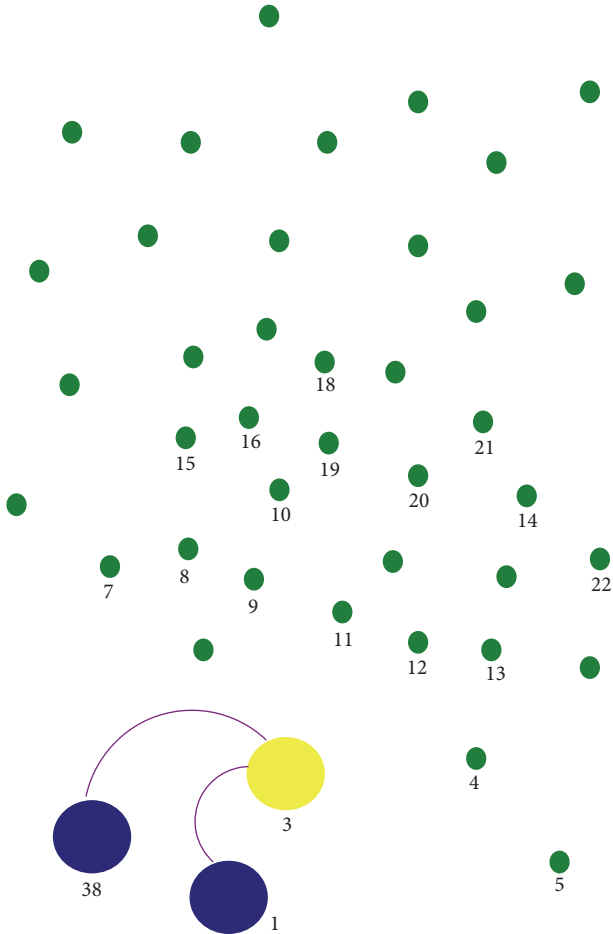


FIGURE 4: Nodes 1 and 38 send information and node 3 receives it.

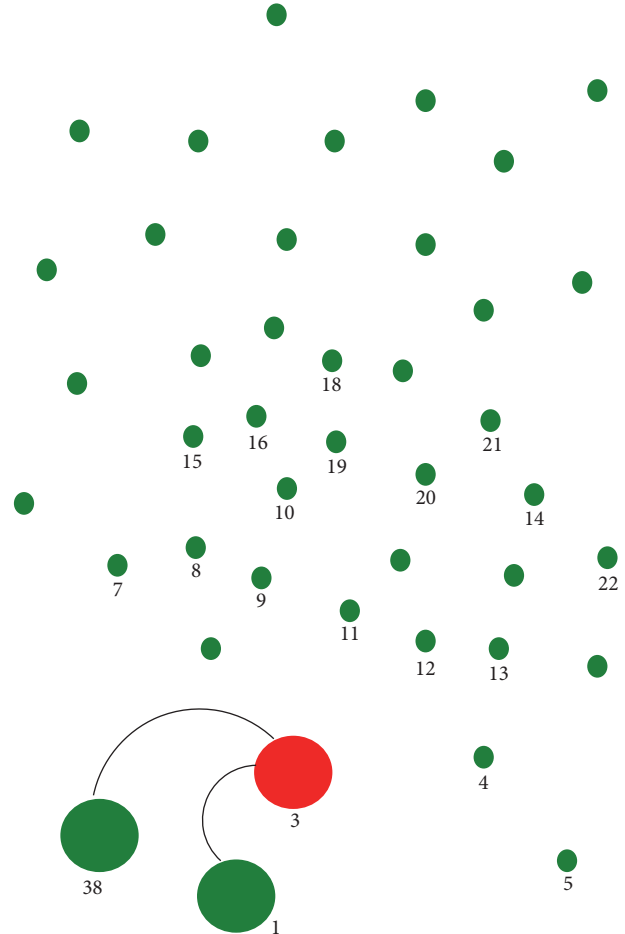


FIGURE 5: Nodes 1 and 38 are idle and node 3 processes information.

used, to determine both the duration of information processing of each node and the delay time of each connection between two nodes.

- (ii) There are 45,145 degrees of the complex network (90,290, input and output degrees) (see Figure 8). This means that the node repository has the greatest number of connections to other nodes given that it is where all the information of the project is stored. The rest of the nodes most relevant to the complex network are the following: develop project management plan, develop schedule, identify risks, and plan procurement management. Therefore, it can be concluded that these nodes are the most likely to be updated or modified. We later define these nodes as strong nodes, because they connect with the greatest number of other nodes.

The nodes with greater degree are the following.

- (i) Node repository (0), degrees: 12,194
- (ii) Node developing project management plan (4), degrees: 6,194
- (iii) Node developing schedule (20), degrees: 5,908
- (iv) Node identifying risks (37), degrees: 4,981

- (v) Node planning procurement management (42), degrees: 4,539.

The nodes with the smallest degree are the following.

- (i) Node planning scope management (9), degrees: 313
- (ii) Node planning cost management (22), degrees: 312
- (iii) Node planning schedule management (15), degrees: 323
- (iv) Node planning risk management (36), degrees: 344
- (v) Node planning risk responses (40), degrees: 353
- (vi) Node of qualitative analysis of risks (38), degrees: 354
- (vii) Node of quality control (28), degrees: 450.

Therefore, the node planning scope management has the least number of connections to other nodes. In addition, these nodes, where only the baselines of the project are determined (schedule, costs, and risks), are less likely to be modified/updated. Thus, they are less important than other nodes are for project management, and thus we define them as weak nodes in project management.

Below, the nodes with the greatest degrees in the complex network simulation are displayed (see Figures 9, 10, and 11).

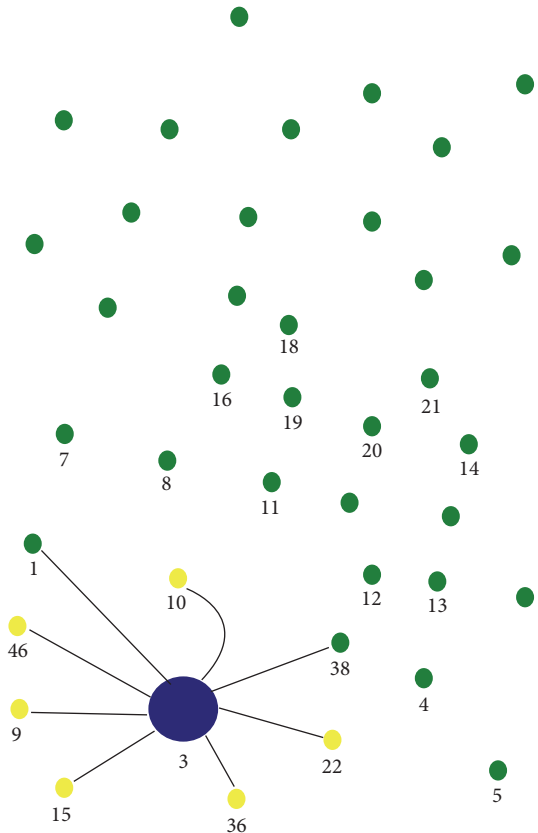


FIGURE 6: Node 3 sends information to other nodes.

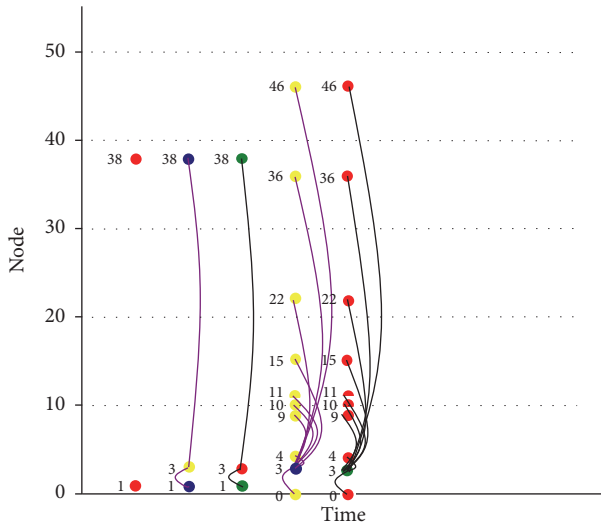


FIGURE 7: Temporal complex network. The dynamics of nodes and connections are depicted. See the text for the definition of the colors and axes.

The planning phase starts at the beginning, the executing phase starts after 21000 time units, the monitoring and controlling phase starts when the number of time units is 30000.

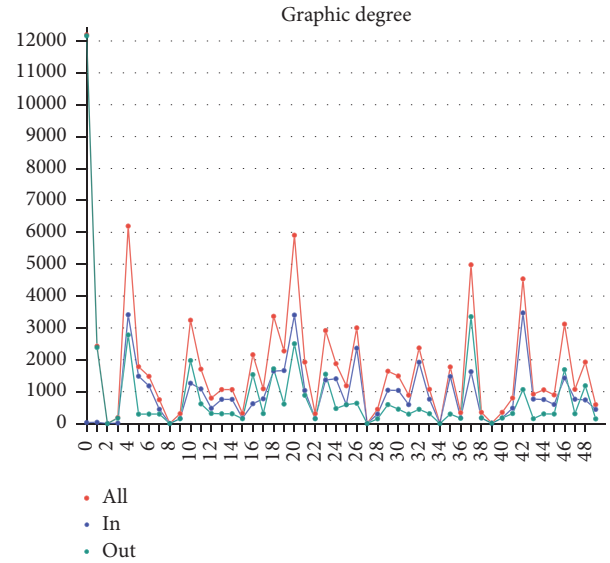


FIGURE 8: Degrees of the 49 nodes; the result of a full simulation. Degrees, input degree, and output degrees are depicted as a result of the simulation. The degrees are colored red, the input degree is colored blue, and the output degree is colored green.

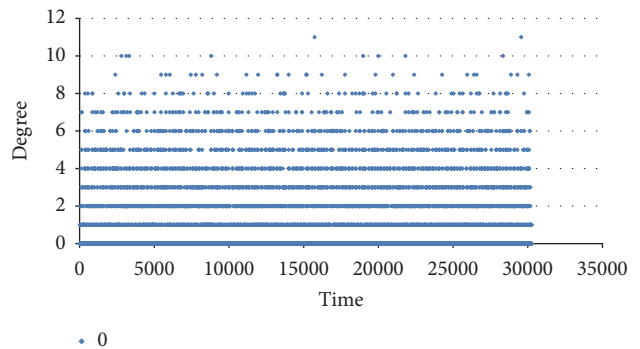


FIGURE 9: Degree of node 0 (repository), throughout the simulation.

In Figure 9 corresponding to node 0 (repository), there is the same pattern of behavior of the degrees in this node throughout the five phases (planning, executing, monitoring and controlling, and closing). However, some peaks are noticeable during the planning phase.

In Figure 10 corresponding to node 20 (developing schedule), the behavior of the degrees of this node is high during the planning and execution phases. Thus, this node is likely to be modified/updated in these two phases but is not as likely to do so in the monitoring and controlling and closing phases.

Using the results of the simulation of all nodes, produced by the software developed for that goal, it is concluded that node 4 (developing project management plan) is the second highest degree node on average and is the node most vulnerable to changes/updates, given the high rate of modifications to it throughout the simulation. Therefore, it is a strong node in the network. Based on the results of the computation of node degrees, strong nodes and weak nodes

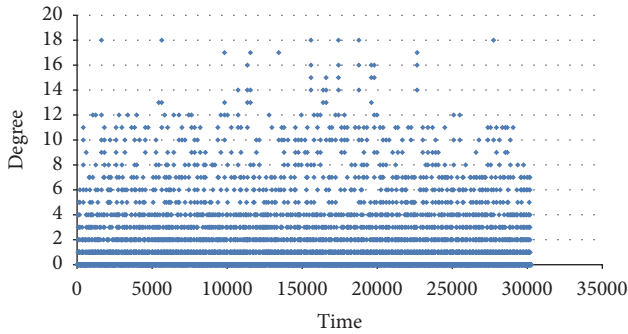


FIGURE 10: Degree of node 20 (developing schedule), throughout the simulation.

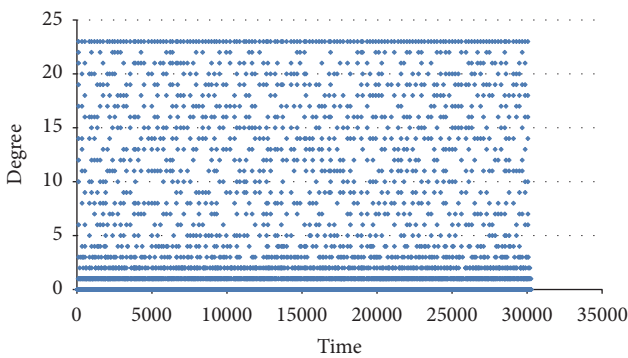


FIGURE 11: Degree of node 4 (developing project management plan), throughout the simulation.

can be identified in the complex network as a product of the project management simulation.

Strong nodes are defined as those with a higher degree during the simulation could be more vulnerable to modifications and have a higher degree of connections with other nodes. In the previous example, the strong nodes are 0, 4, 20, 7, and 42.

On the other hand, weak nodes are defined as those with a smaller degree during the simulation, those that could be less vulnerable to modifications, and those that have a lower degree of connections with other nodes. In the previous example, the weak nodes are 9, 22, 15, 36, 40, 38, and 28.

It can be inferred that if a strong node is not activated (e.g., from a lack of information), the project could be in a greater state of uncertainty.

The definition of strong and weak nodes, and their potential activation, is essential and new in complex project management given the elucidation of critical nodes. These critical nodes require more attention to succeed in projects of great complexity.

The following are other important measurements of the complex network.

- (i) Assortativity: $-0,01052$: If the value is <0 , then the relationships in the network are established between nodes of different degree. In project management, this is because a node can in turn update/modify

a secondary network, and nodes with higher degree interact with nodes with lower grade.

- (ii) Density: 18.42653: Such a high density is noteworthy, given that it could indicate a high effect of new information in the system because there are nodes connected to other nodes, those ones connected to others, and so on. Thus, updates produced by new or modified information add complexity to project management.
- (iii) Diameter: 3: it is the maximum distance between two nodes in the network. In project management, the diameter measurement yields a measure of how far away the nodes can be.
- (iv) Adjacency: the nodes with the greatest adjacency are the following.

- (a) Node 42 (plan procurement management) with node 37 (identify risks): adjacency value 1.055
- (b) Node 42 (plan procurement management) with node 0 (repository): adjacency value 903
- (c) Node 42 (plan procurement management) with node 10 (collect requirements): adjacency value 899
- (d) Node 20 (develop schedule) with node 4 (develop project management plan): adjacency value 776
- (e) Node 20 (develop schedule) with node 0 (repository): adjacency value 775
- (f) Node 19 (develop project management plan) with node 0 (repository): adjacency value 602.

Nodes with higher adjacency are those that are in groups of nodes with higher degree. In this case, they are node 42 (plan procurement management) and node 20 (develop schedule). These nodes in the project management are strong nodes that generate connections with a high number of other nodes, being strong nodes in the complex network.

- (i) Betweenness: nodes with the highest values are the following: node 0 (repository): 809.949, node 4 (develop project management plan): 486.7958, and node 20 (develop schedule): 123.04543. These are strong nodes through which the most amount of information passes, with the most amount of control over the network.
- (ii) Closeness: node 2 (novelties): 0.02; node 8 (close the project or phase): 0.342657343: For the analysis, node 2 can be ignored because throughout the simulation the possibility of novelties was not included. Thus, node 8 is closest to the center of the network. The node closest to the center of the network can be considered that where the simulation ends, because either a phase of the project ends there, or the entire project ends.
- (iii) Clustering: node 1 (repository) is the most connected node throughout the simulation, given that it is where all the information is stored.

4. Discussion and Conclusions

The classic perspective is not sufficient to understand the high number of interrelations established in the different phases of a complex project. Thus, project management should be approached from the perspective of complexity, and a complex project should be understood as a complex system. Using the algorithmic methodology established in this study, a modeling scheme can be constructed for any type of project. This is the main contribution of this work, and the stochastic simulation and subsequent analysis are one of the fundamental results of this research.

Weak nodes and strong nodes can be identified, thereby identifying the most vulnerable nodes to be modified/updated and those that, if not activated, could generate higher thresholds of uncertainty for a project. The strong nodes in the standard structure of PMBOK are node 4 (develop project management plan) and node 20 (develop schedule). The weak nodes are node 9 (plan scope management) and node 22 (plan cost management).

The strong nodes in the standard structure of PMBOK are node 4 (develop project management plan) and node 20 (develop schedule). The weak nodes are node 9 (plan scope management) and node 22 (plan cost management).

These results suggest that, given PMBOK, the scope and costs should have almost no changes/updates throughout the management of the project. In addition, the classic perspective assumes that behavior is deterministic from planning to closing of the project or phase. Thus, some rigidity can be established in the nature of the project when it is in a complex situation.

Given this, the following important question arises. What happens if nodes 4 and 20 disappear from the network (e.g., if they fail to be activated because of a lack of information)? To answer this question, new simulations are run in which nodes 4 (develop project management plan) and 20 (develop schedule) are not activated in the modeling of the network.

The results are as follows.

- (i) Degrees: 29,552, with a 32% decrease from the initial simulation (degrees: 90,290): This confirms that the mentioned nodes have a significant influence over the others and that the information outputted from such is important for the project management. Higher levels of uncertainty are generated by the absence of the nodes, which is key for the project.
- (ii) Density: it is 6.031020, with a 32% decrease from the initial simulation (density: 18.42653).
- (iii) Diameter becomes 4, increasing the maximum distance between two nodes in the network. This suggests that the nodes (4 and 20) are bridges that aid in the flow of information throughout the network.

Another conclusion is that increasing importance should be attributed to the use of tools based on the science of complexity in order to interpret complex project management, given that they provide additional information not available from classic-perspective tools.

Finally, the structure of the process network described can be used in the future to advance different types of simulations for any type of project. In particular, any sequence in the construction of the network can be simulated. As new information is delivered, the network updates the nodes/sub-processes as required.

This study establishes measurements of uncertainty, efficiency, and robustness, based on the complex network, which can be interpreted for the management of complex projects. This enables decision making prior to the start of the execution of a project. The report of this research will be described in a future article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors thank Professor Rafael Rentería for his contributions to this article and his important reflections on complex projects. Gerard Olivar-Tost acknowledges DIMA Project Grant *Modelamiento Avanzado de Mercados de Energía Eléctrica para Toma de Decisiones de Inversión y Establecimiento de Políticas Code 35467* from Universidad Nacional de Colombia.

References

- [1] S. J. Whitty and H. Maylor, "And then came complex project management (revised)," *International Journal of Project Management*, vol. 27, no. 3, pp. 304–310, 2009.
- [2] P. W. G. Morris, J. R. Pinto, and J. Söderlund, "The oxford handbook of project management," 2012.
- [3] M. Padalkar and S. Gopinath, "Are complexity and uncertainty distinct concepts in project management? a taxonomical examination from literatura," *International Journal of Project Management*, 2016.
- [4] R. Kolisch, "Serial and parallel resource-constrained project scheduling methods revisited: theory and computation," *European Journal of Operational Research*, vol. 90, no. 2, pp. 320–333, 1996.
- [5] W. Herroelen, B. De Reyck, and E. Demeulemeester, "Resource-constrained project scheduling: a survey of recent developments," *Computers & Operations Research*, vol. 25, no. 4, pp. 279–302, 1998.
- [6] R. Kolisch and R. Padman, "An integrated survey of deterministic project scheduling," *Omega*, vol. 29, no. 3, pp. 249–272, 2001.
- [7] W. Herroelen and R. Leus, "Robust and reactive project scheduling: a review and classification of procedures," *International Journal of Production Research*, vol. 42, no. 8, pp. 1599–1620, 2004.
- [8] S. Hartmann and D. Briskorn, "A survey of variants and extensions of the resource-constrained project scheduling problem," *European Journal of Operational Research*, vol. 207, no. 1, pp. 1–14, 2010.
- [9] C. S. Lim and M. Z. Mohamed, "Criteria of project success: an explanatory re-examination," *International Journal of Project Management*, vol. 21, pp. 411–418, 1999.

- [10] D. Baccarini, "The Logical Framework Method for Defining Project Success," *Project Management Journal*, vol. 30, no. 4, pp. 25–32, 1999.
- [11] T. J. Kloppenborg and W. A. Opfer, "The current state of project management research: trends, interpretations, and predictions," *Project Management Journal*, vol. 33, no. 2, pp. 5–18, 2002.
- [12] A. S. Pillai, A. Joshi, and K. S. Rao, "Performance measurement of R&D projects in a multi-project, concurrent engineering environment," *International Journal of Project Management*, vol. 20, no. 2, pp. 165–177.
- [13] D. Tesch, T. J. Kloppenborg, and J. K. Stemmer, "Project management learning: what the literature has to say," *Project Management Journal*, vol. 34, no. 4, pp. 33–39, 2003.
- [14] J. R. Turner and R. Müller, "The project manager's leadership style as a success factor on projects: a literature review," *Project Management Journal*, vol. 36, no. 1, pp. 49–61, 2005.
- [15] K. Jugdev and R. Müller, "A retrospective look at our evolving understanding of project success," *Project Management Journal*, vol. 36, no. 4, pp. 19–31, 2005.
- [16] S. Rozenes, G. Vitner, and S. Spraggett, *Project control: literature review*, Project Management Institute, 2006.
- [17] M. Huemann, A. Keegan, and J. R. Turner, "Human resource management in the project-oriented company: a review," *International Journal of Project Management*, vol. 25, no. 3, pp. 315–323, 2007.
- [18] G. P. Prabhakar, "What is project success: a literature review," *International Journal of Business and Management*, vol. 3, no. 9, 2009.
- [19] L. A. Ika, "Project success as a topic in project management journals," *Project Management Journal*, vol. 40, no. 4, pp. 6–19, 2009.
- [20] C. Barclay and K.-M. Osei-Bryson, "Project performance development framework: an approach for developing performance criteria & measures for information systems (IS) projects," *International Journal of Production Economics*, vol. 124, no. 1, pp. 272–292, 2010.
- [21] H. Wi and M. Jung, "Modeling and analysis of project performance factors in an extended project-oriented virtual organization (EProVO)," *Expert Systems with Applications*, vol. 37, no. 2, pp. 1143–1151, 2010.
- [22] R. Müller and K. Jugdev, "Critical success factors in projects: Pinto, Slevin, and Prescott – the elucidation of project success," *International Journal of Managing Projects in Business*, vol. 5, no. 4, pp. 757–775, 2012.
- [23] A. de Wit, "Measurement of project success," *International Journal of Project Management*, vol. 6, no. 3, pp. 164–170, 1988.
- [24] T. Cooke-Davies, "The 'real' success factors on projects," *International Journal of Project Management*, vol. 20, no. 3, pp. 185–190, 2002.
- [25] A. Belout and C. Gauvreau, "Factors influencing project success: the impact of human resource management," *International Journal of Project Management*, vol. 22, no. 1, pp. 1–11, 2004.
- [26] D. Aloini, R. Dulmin, and V. Mininno, "Risk management in ERP project introduction: review of the literature," *Information and Management*, vol. 44, no. 6, pp. 547–567, 2007.
- [27] P. Littau, N. J. Jujagiri, and G. Adlbrecht, "25 Years of stakeholder theory in project management literature (1984–2009)," *Project Management Journal*, vol. 41, no. 4, pp. 17–29, 2010.
- [28] M. Padalkar and S. Gopinath, "Delays in projects: a game-theoretic study," in *Trends and Research in the Decision Sciences: Best Papers from the 2014 Annual Conference*, M. Warkentin, Ed., pp. 191–212, Pearson Education, Upper Saddle River, NJ, USA.
- [29] F. E. Grubbs, "Letter to the Editor—Attempts to validate certain PERT statistics or 'Picking on PERT,'" *Operations Research*, vol. 10, no. 6, pp. 912–915, 1962.
- [30] K. R. MacCrimmon and C. A. Ryavec, "An analytical study of the PERT assumptions," *Operations Research*, vol. 12, no. 1, pp. 16–37, 1964.
- [31] R. J. Schonberger, "Why projects are 'always' late: a rationale based on manual simulation of a PERT/CPM network," *Interfaces*, vol. 11, no. 5, pp. 66–70, 1981.
- [32] A. J. Shenhar and D. Dvir, *Reinventing Project Management*, Harvard Business School Press, 2007.
- [33] J. M. Burt, "Planning and dynamic control of projects under uncertainty," *Management Science*, vol. 24, no. 3, pp. 249–258, 1977.
- [34] T. M. Cook and R. H. Jennings, "Estimating a project's completion time distribution using intelligent simulation methods," *Journal of the Operational Research Society*, vol. 30, no. 12, pp. 1103–1108, 1979.
- [35] T. M. Williams, "Criticality in stochastic networks," *Journal of the Operational Research Society*, vol. 43, no. 4, pp. 353–357, 1992.
- [36] R. A. Bowman, "Efficient estimation of arc criticalities in stochastic activity networks," *Management Science*, vol. 41, no. 1, pp. 58–67, 1995.
- [37] J. G. Cho and B. J. Yum, "An uncertainty importance measure of activities in PERT networks," *International Journal of Production Research*, vol. 35, no. 10, pp. 2737–2758, 1997.
- [38] S. E. Elmaghraby, Y. Fathi, and M. R. Taner, "On the sensitivity of project variability to activity mean duration," *International Journal of Production Economics*, vol. 62, no. 3, pp. 219–232, 1999.
- [39] C. Chapman and S. Ward, "Estimation and evaluation of uncertainty: a minimalist first pass approach," *International Journal of Project Management*, vol. 18, no. 6, pp. 369–383, 2000.
- [40] T. Williams, C. Eden, F. Ackerman, and A. Tait, "Vicious circles of parallelism," *International Journal of Project Management*, vol. 13, no. 3, pp. 151–155, 1995.
- [41] A. G. Rodrigues and T. M. Williams, "System dynamics in project management: assessing the impacts of client behaviour on project performance," *Journal of the Operational Research Society*, pp. 2–15, 1998.
- [42] T. M. Williams, "The need for new paradigms for complex projects," *International Journal of Project Management*, vol. 17, no. 5, pp. 269–273, 1999.
- [43] C. Eden, T. Williams, F. Ackermann, and S. Howick, "The role of feedback dynamics in disruption and delay on the nature of disruption and delay (D and D) in major projects," *Journal of the Operational Research Society*, vol. 51, no. 3, pp. 291–300, 2000.
- [44] S. Chanas and P. Zieliński, "Critical path analysis in the network with fuzzy activity times," *Fuzzy Sets and Systems*, vol. 122, no. 2, pp. 195–204, 2001.
- [45] T. R. Browning and S. D. Eppinger, "Modeling impacts of process architecture on cost and schedule risk in product development," *IEEE Transactions on Engineering Management*, vol. 49, no. 4, pp. 428–442, 2002.
- [46] S. van de Vonder, E. Demeulemeester, W. Herroelen, and R. Leus, "The use of buffers in project management: the trade-off between stability and makespan," *International Journal of Production Economics*, vol. 97, no. 2, pp. 227–240, 2005.

- [47] C. Jensen, S. Johansson, and M. Löfström, "Project relationships—a model for analyzing interactional uncertainty," *International Journal of Project Management*, vol. 24, no. 1, pp. 4–12, 2006.
- [48] S.-P. Chen, "Analysis of critical paths in a project network with fuzzy activity times," *European Journal of Operational Research*, vol. 183, no. 1, pp. 442–459, 2007.
- [49] A. J. Shenhar and D. Dvir, "Toward a typological theory of project management," *Research Policy*, vol. 25, no. 4, pp. 607–632, 1996.
- [50] A. J. Shenhar, "One size does not fit all projects: exploring classical contingency domains," *Management Science*, vol. 47, no. 3, pp. 394–414, 2001.
- [51] J. Söderlund, "Building theories of project management: past research, questions for the future," *International Journal of Project Management*, vol. 22, no. 3, pp. 183–191, 2004.
- [52] S. Cicmil, T. Williams, J. Thomas, and D. Hodgson, "Rethinking project management: researching the actuality of projects," *International Journal of Project Management*, vol. 24, no. 8, pp. 675–686, 2006.
- [53] H. J. Smyth and P. W. Morris, "An epistemological evaluation of research into projects and their management: methodological issues," *International Journal of Project Management*, vol. 25, no. 4, pp. 423–436, 2007.
- [54] P. W. G. Morris, "Research and the future of project management," *International Journal of Managing Projects in Business*, vol. 3, no. 1, pp. 139–146, 2010.
- [55] M. Jacobsson and A. Söderholm, "Breaking out of the strait-jacket of project research: in search of contribution," *International Journal of Managing Projects in Business*, vol. 4, no. 3, pp. 378–388, 2011.
- [56] S. Austin, A. Newton, J. Steele, and P. Waskett, "Modelling and managing project complexity," *International Journal of Project Management*, vol. 20, no. 3, pp. 191–198, 2002.
- [57] S. Howick and C. Eden, "The impact of disruption and delay when compressing large projects: going for incentives?" *Journal of the Operational Research Society*, vol. 52, no. 1, pp. 26–34, 2001.
- [58] W. Xia and G. Lee, "Grasping the complexity of is development projects," *Communications of the ACM*, vol. 47, no. 5, pp. 68–74, 2004.
- [59] W. Xia and G. Lee, "Complexity of information systems development projects: conceptualization and measurement development," *Journal of Management Information Systems*, vol. 22, no. 1, pp. 45–83, 2005.
- [60] S. H. Cho and S. D. Eppinger, "A simulation-based process model for managing complex design projects," *IEEE Transactions on Engineering Management*, vol. 52, no. 3, pp. 316–328, 2005.
- [61] M. Danilovic and T. R. Browning, "Managing complex product development projects with design structure matrices and domain mapping matrices," *International Journal of Project Management*, vol. 25, no. 3, pp. 300–314, 2007.
- [62] D. Baccarini, "The concept of project complexity—a review," *International Journal of Project Management*, vol. 14, no. 4, pp. 201–204, 1996.
- [63] M. T. Pich, C. H. Loch, and A. De Meyer, "On uncertainty, ambiguity, and complexity in project management," *Management Science*, vol. 48, no. 8, pp. 1008–1023, 2002.
- [64] S. C. Sommer and C. H. Loch, "Selectionism and learning in projects with complexity and unforeseeable uncertainty," *Management Science*, vol. 50, no. 10, pp. 1334–1347, 2004.
- [65] H. Benbya and B. McKelvey, "Toward a complexity theory of information systems development," *Information Technology and People*, vol. 19, no. 1, pp. 12–34, 2006.
- [66] T. Cooke-Davies, S. Cicmil, L. Crawford, and K. Richardson, "Mapping the strange landscape of complexity theory, and its relationship to project management," *Project Management Journal*, vol. 38, no. 2, pp. 50–61, 2007.
- [67] J. Geraldi and G. Adlbrecht, "On faith, fact, and interaction in projects," *Project Management Journal*, vol. 38, no. 1, pp. 32–43, 2007.
- [68] H. Maylor, R. Vidgen, and S. Carver, "Managerial complexity in project based operations: a grounded model and its implications for practice," *Project Management Journal*, vol. 39, no. S1, pp. S15–S26, 2008.
- [69] L.-A. Vidal and F. Marle, "Understanding project complexity: implications on project management," *Kybernetes*, vol. 37, no. 8, pp. 1094–1110, 2008.
- [70] Girmscheid and Brockmann, "The inherent complexity of large scale engineering projects," in *The Annual Publication of International Project Management Association*, vol. 29, pp. 22–26, 2008.
- [71] T. Brady and A. Davies, "From hero to hubris? reconsidering the project management of Heathrows Terminal 5," *International Journal of Project Management*, vol. 28, no. 2, pp. 151–157, 2010.
- [72] S. Lenfle, "The strategy of parallel approaches in projects with unforeseeable uncertainty: the Manhattan case in retrospect," *International Journal of Project Management*, vol. 29, no. 4, pp. 359–373, 2011.
- [73] R. V. Ramasesh and T. R. Browning, "A conceptual framework for tackling knowable unknown unknowns in project management," *Journal of Operations Management*, vol. 32, no. 4, pp. 190–204, 2014.
- [74] M. V. Tatikonda and S. R. Rosenthal, "Technology novelty, project complexity, and product development project execution success: a deeper look at task uncertainty in product innovation," *IEEE Transactions on Engineering Management*, vol. 47, no. 1, pp. 74–87, 2000.
- [75] J. R. Turner and R. Müller, "On the nature of the project as a temporary organization," *International Journal of Project Management*, vol. 21, no. 1, pp. 1–8, 2003.
- [76] S. Ward and C. Chapman, "Transforming project risk management into project uncertainty management," *International Journal of Project Management*, vol. 21, no. 2, pp. 97–105, 2003.
- [77] R. Atkinson, L. Crawford, and S. Ward, "Fundamental uncertainties in projects and the scope of project management," *International Journal of Project Management*, vol. 24, no. 8, pp. 687–698, 2006.
- [78] P. Anderson, "Complexity theory and organization science," *Organization Science*, vol. 10, no. 3, pp. 216–323, 1999.
- [79] K. Remington and J. Pollack, "Tools for complex project," 2010.
- [80] D. Arellano, J. Danti, and M. F. Pérez, *Proyectos y Sistemas*, PMI-INCOSE, 2016, (Spanish).
- [81] R. García, *Sistemas Complejos*, Editorial GEDISA, 2006, (Spanish).
- [82] C. Ruiz-Martin and D. J. Poza, "Project configuration by means of network theory," *International Journal of Project Management*, vol. 33, no. 8, pp. 1755–1767, 2015.
- [83] P. Holme and J. Saramaki, "Temporal networks," *Physics Reports*, vol. 519, no. 3, pp. 97–125, 2012.
- [84] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. W. Hwang, "Complex networks: structure and dynamics," *Physics Reports*, vol. 424, no. 4–5, pp. 175–308, 2006.

- [85] S. Boccaletti, G. Bianconi, R. Criado, C. I. del Genio et al., “The structure and dynamics of multilayer networks,” *Physics Reports*, vol. 544, no. 1, pp. 1–122, 2014.
- [86] H. Sayama, *Introduction to the Modeling and Analysis of Complex Systems*, 2015.
- [87] A. K., *Network Analysis Literacy: A Practical Approach to The Analysis of Networks*, Springer, 2016.

Research Article

A Language as a Self-Organized Critical System

Vasilii A. Gromov and Anastasia M. Migrina

School of Applied Mathematics, Oles Honchar Dnipropetrovsk National University, Gagarina Av. 72, Dnipropetrovsk 49010, Ukraine

Correspondence should be addressed to Vasilii A. Gromov; stroller@rambler.ru

Received 2 May 2017; Revised 3 September 2017; Accepted 31 October 2017; Published 19 November 2017

Academic Editor: Gerard Olivar

Copyright © 2017 Vasilii A. Gromov and Anastasia M. Migrina. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

A natural language (represented by texts generated by native speakers) is considered as a complex system, and the type thereof to which natural languages belong is ascertained. Namely, the authors hypothesize that a language is a self-organized critical system and that the texts of a language are “avalanches” flowing down its word cooccurrence graph. The respective statistical characteristics for distributions of the number of words in the texts of English and Russian languages are calculated; the samples were constructed on the basis of corpora of literary texts and of a set of social media messages (as a substitution to the oral speech). The analysis found that the number of words in the texts obeys power-law distribution.

1. Introduction

Since natural languages gradually came to be regarded as complex systems, a means to study linguistic processes changed from descriptive approaches to formal analysis aiming to construct mathematical model for the operation and the development of language—this change is both a reason and an effect of “the linguistic turn” (the term is due to [1]).

In the frameworks of this approach, it is possible to pursue two main avenues of inquiry: The first one is aimed at constructing a theory of natural languages grammars [2]. The second one is associated with analysis of language statistical characteristics: Primary emphasis is placed, from seminal studies by Zipf [3], on distributions for separate words for written language [4–8] and for various graphs reflecting language features [9]. The principal result of these studies is a class of distributions describing natural language features [10]; the class comprises power (heavy-tail) distributions with various values of exponents: power laws manifest themselves in word’s frequency in language [3], in syntactic networks [5–7, 11], in frequency of letter sequences in vocabularies [8], and so forth.

Meanwhile, a shift in XX century philosophy (akin to the Copernican revolution in astronomy) suggests considering a

language as an integral whole: “a man is just a place where a language speaks itself” [12]. Therefore, since above all else a language is a unified communicative tool to convey the meanings and its text (either a literary work or a tweet) is usually a meaningful, complete message, a text becomes a basic unit for such analysis.

The authors hypothesize that a natural language is a self-organized critical system (SOC) [13, 14] and the texts of a language are “avalanches” (as those are defined by Bak) flowing down the word cooccurrence graph of the respective language; large avalanches correspond to literary works, while smaller ones are associated with messages from social media. It is worth noting that a self-organized critical system conventionally features [13, 14]

- (1) a space of elements able to be in two states, active and passive, along with a set of rules to describe how a change of state of an element affects states of the other ones,
- (2) the “avalanches” in the space that are chain reactions of elements’ state changes triggered by changes of other elements,
- (3) a power-law distribution governing avalanche sizes.

For a language system, a semantic space plays a part of the space at issue, and the rules are reduced to syntactic

and semantic rules of the respective language. In the present study the space was formalized as a cooccurrence graph—the authors are aware that one-to-one correspondence between semantic space and vocabulary is absent, but they assume that the latter approximates the former somehow. Vertices of the graph correspond to words and an edge is present if and only if the words associated with its incident vertices occur simultaneously in the same text of the sample involved, once or more. As indicated above, an avalanche, in this context, is a text of a language, and the hypothesis that sizes of avalanches obey a power-law distribution forms the subject of the present study. One should mention that real-world (unfolding over time) SOC-systems usually exhibit long periods of slow evolution as opposed to short periods of fast evolution when the system space is changed drastically; similar phenomenon is reported to take place for evolving language systems.

Another point of interest here is the emergence of a gigantic volume of information reflecting, in essence, spoken language (parole as opposed to langue according to Saussurean terminology [15], (second) Orality versus Literacy [16]) that is texts posted by users in social networks of every sort and kind (Facebook, Reddit, and so on); this makes it possible to explore this domain of human communication. In this context, the primary problem is to compare statistical characteristics calculated by means of corpus of literary texts, on the one hand, and by means of a set of texts written by social networks users. If these characteristics appear to be statistically equal, this may give proof to the idea of language unity as a complex system; and if so, written and spoken language are merely different projections of internal dynamics of this complex system (synchronic unity of language). On the other hand, the characteristic at issue calculated for different time periods (one is to be restricted in this case to the analysis of written language) can be compared in order to verify unity of the linguistic system unfolding in time (diachronic unity of language). The present work is focused on the study (and comparison) of statistical characteristics of texts sets for English and Russian languages being considered in their synchronic and diachronic aspects.

The first paper to be cited among recent studies of distributions (mainly power-law distribution) observed for oral and written language is a brilliant review [17]. Here, the object of study is a particular text, and its basic unit is a word or a sentence—distributions of characteristics of these objects form the subjects of the overwhelming majority of papers for this line of investigation [18]. For example, Font-Clos et al. [19] examine dependence text length versus statistical properties of word occurrences; the authors reveal that the distribution obeys the power and investigate its relationship with Zipf's and Heaps' (Herdan's) laws (the former states that the vocabulary grows as a power function of text's length) [20].

Another object of study that generates power-law distributions is a representation of language semantic and syntactic relations using various discrete structures: semantic nets [17], global syntactic dependency trees [21, 22], cooccurrence graphs [18], and others. Both conventional methods aimed at exploring these structures as complex networks [23]

and random walks on these structures result in power-law distributions [17, 21, 22, 24, 25]. A basic unit is a word or a sentence likewise.

The subject of the present paper is language as a whole; texts (semantic “avalanches”) are considered as its basic units. The rest of the paper is organized as follows. The next section outlines methods used to estimate distribution parameters; the third provides results for both English and Russian languages. The fourth section discusses results; finally, the last section presents conclusions.

2. Methods

The choice of the languages is determined, apart from the availability of voluminous corpora of texts (for written and spoken languages) for them, by their qualitative difference in grammar structure: Russian is an inflected language, while in English inflections are rather rare; Russian is characterized by flexible word order in a sentence, whereas word order in English language is strict, and rare exceptions are constrained by stringent rules [26, 27].

We used two different approaches to test statistical hypothesis in question. The first one utilizing the concept of data collapse is considered in greater detail in the monograph by Pruessner [14]; the second one (grounded on the Kolmogorov-Smirnov (KS) criterion) is proposed in the source [28]. Both methods not only evaluate the exponents but also cut off the smallest elements of the sample that usually do not fit a power-law. The first method also cuts off the largest nonfitting elements. It is worth noting that this phenomenon (the sharp distinction between the largest [smallest] elements and all others) seems to be a salient characteristic of real-world data following power-law distributions [13, 14].

The first approach we dwell on briefly is that using the concept of data collapse [14]. It assumes that the distribution generating the data has the following probability density function:

$$p(x) = ax^{-\tau} g\left(\frac{x}{bL^D}\right) \quad (1)$$

with x which can be a continuous or discrete random variable, metric factors a and b , characteristic dimension of a system L , scaling function g , and scaling exponents τ and D as well. The distribution follows the modified power-law within the interval bounded by the lower and upper cutoffs x_{\min} and x_{\max} . The scaling function g (that distinguishes the distribution from a canonical power-law) fits many real-world systems obeying heavy-tail distributions [14]. The quantity bL^D determines the characteristic upper cutoff.

The raw data is taken to be previously binned, that is, grouped together and averaged over the observations belonging to the same group; we used the exponential binning for its suitability for this kind of data [14].

If the null hypothesis (that the sample under study is generated from the distribution equation (1)) is true, then $\bar{p}(x)x^{-\tau}$ plotted against $\xi = x/x_c(L)$, where $x_c(L) = bL^D$, gives the same function $ag(\xi)$ for various L , where $\bar{p}(x)$ is the empirical probability density function, and τ is the true

value of the power exponent. The phenomenon is given the title of data collapse. Thus, for given data generated from the power-law distribution, $\bar{p}(x)x^{-\tau}$ (as a function of ξ) plotted for various L is superimposed on each other.

Therefore, the respective goodness-of-fit test for power-law distribution involves the following steps:

- (1) binning of the raw data
- (2) plotting $\bar{p}(x)x^{-\tau}$ against $\xi = x/x_c(L)$ for various L using “apparent exponent” $\hat{\tau}$ (rough estimate of τ)—such a plot comprises a nonhorizontal straight line and a characteristic nonlinear curve, whose extremum is called a landmark (x_c represents its coordinate)
- (3) the refinement of the value with the employment of the least squares method applied to the landmarks.

As a result, the plots merge into a single horizontal straight line for the section of the domain of definition for which power-law holds true (between the lower and upper cutoffs).

The second approach [28] is applicable to power-law distribution without scaling function:

$$p(x) = \begin{cases} 0, & x \leq x_{\min} \\ ax^{-\tau}, & x_{\min} < x \end{cases} \quad (2)$$

with normalization constant $a = 1/\zeta(\tau, x_{\min})$, where

$$\zeta(\tau, x_{\min}) = \sum_{i=1}^n (i + x_{\min})^{-\tau} \quad (3)$$

is the generalized (Hurwitz) zeta function. The method implies that all practicable values of the lower cutoff x_{\min} are considered; for each x_{\min} the estimate of the power exponent (with the maximum likelihood principle in mind) is calculated from

$$\hat{\tau} = 1 + n \left[\sum_{i=1}^n \ln \frac{x_i}{x_{\min}} \right]^{-1}; \quad (4)$$

see [28]; for each x_{\min} the Kolmogorov-Smirnov statistic $S = \sup_{x \geq x_{\min}} |P(x) - \bar{P}(x)|$ (where $P(x)$ is the cumulative distribution function (CDF) with estimated value of τ and $\bar{P}(x)$ is the empirical CDF) is calculated. The eventual estimate of x_{\min} minimizes $S(x_{\min})$. For real-world data, the function $S(x_{\min})$ possesses, usually, several local minima; it is often reasonable not to choose a global minimum but the local minimum closest to x_{\min} , the lower boundary of the domain of definition, provided a value of the statistic at it does not differ significantly from that at a global minimum.

3. Power-Law Distributions for the English and Russian Languages

To test the null hypothesis in question (that the number of words in the texts obeys a power-law distribution; (1) is

used to verify data collapse, while the method based on the KS statistics employs (2), (3)), two samples were generated on the basis of corpora of literary texts for these languages and of a set of Reddit messages (or its Russian counterpart Pikabu). The resulting samples sizes for English language are 9820 (literary works), 5016 (Reddit), and 14836 (joint sample); for Russian language they are 12683 (literary works), 6005 (Pikabu), and 18688 (joint sample). For the method based on the concept of data collapse, the size of vocabulary used to generate texts serves as characteristic dimension of a system L . To obtain samples for various L , one resamples down the initial sample deleting randomly $\lambda\%$ of words from the complete vocabulary and then from all the texts used. This brings about the generation of new samples corresponding to the characteristic dimension of $L(1 - \lambda/100)$.

Figure 1 presents a dependence of the number of words in a text in double-logarithmic scale on a rank of the text in the sample; namely, Figures 1(a), 1(b), and 1(c) correspond to the joint sample, to the sample constructed on the basis of literary works, and to poetry works for the English language, respectively; Figures 1(d), 1(e), and 1(f) exhibit the same dependence for the Russian, respectively; a dashed straight line in each subfigure corresponds to power distribution with an exponent estimated using data collapse.

Figure 2 (in the coordinates $(x^{-\tau}\bar{p}(x), \xi = x/x_c)$, x_c is a landmark coordinate) shows data collapse for the joint samples for English (Figure 2(a)) and Russian (Figure 2(d)) languages. Raw data was binned exponentially with bin sizes $B_j = \lfloor cR^j \rfloor (R-1)$ ($c = 1$ and $R = 1.4$). Red colour (with discs) stands for the complete vocabulary (of size L), grey colour (with squares) is for the vocabulary of size $0.9L$, blue colour (with diamonds) is for $0.8L$, black colour (with triangles) is for $0.7L$, orange colour (with upturned triangles) is for $0.6L$, and, finally, purple colour (with circles) is for $0.5L$; the curves are dragged apart a little bit in order to make it possible to distinguish as they are superimposed owing to data collapse. Figures 2(b) and 2(e) present the same dependence for samples constructed using corpora of literary texts for English (Figure 2(b)) and Russian (Figure 2(e)) languages. Figures 2(c) and 2(f) demonstrate data collapse for poetry samples for English and Russian languages, respectively.

The results obtained using both approaches are presented in Table 1; the table includes results for samples constructed on the basis of literary works and of messages of social media and for joint sample as well for both languages. Each cell contains estimates for a power exponent and (in parentheses) for a lower cutoff x_{\min} . The above results suggest synchronic unity for both languages because of a good agreement of estimates for power exponents and lower cutoffs calculated for literary works and for social media messages.

In order to deal with the problem of diachronic unity of a language, authors confined themselves to the samples generated on the basis of literary works created before the 20th century and in the 20th century for both the English and Russian languages (respective samples sizes amount to 7179 and 2641 for the English language, 5758 and 6925 for the Russian language). Table 2 exhibits the respective results.

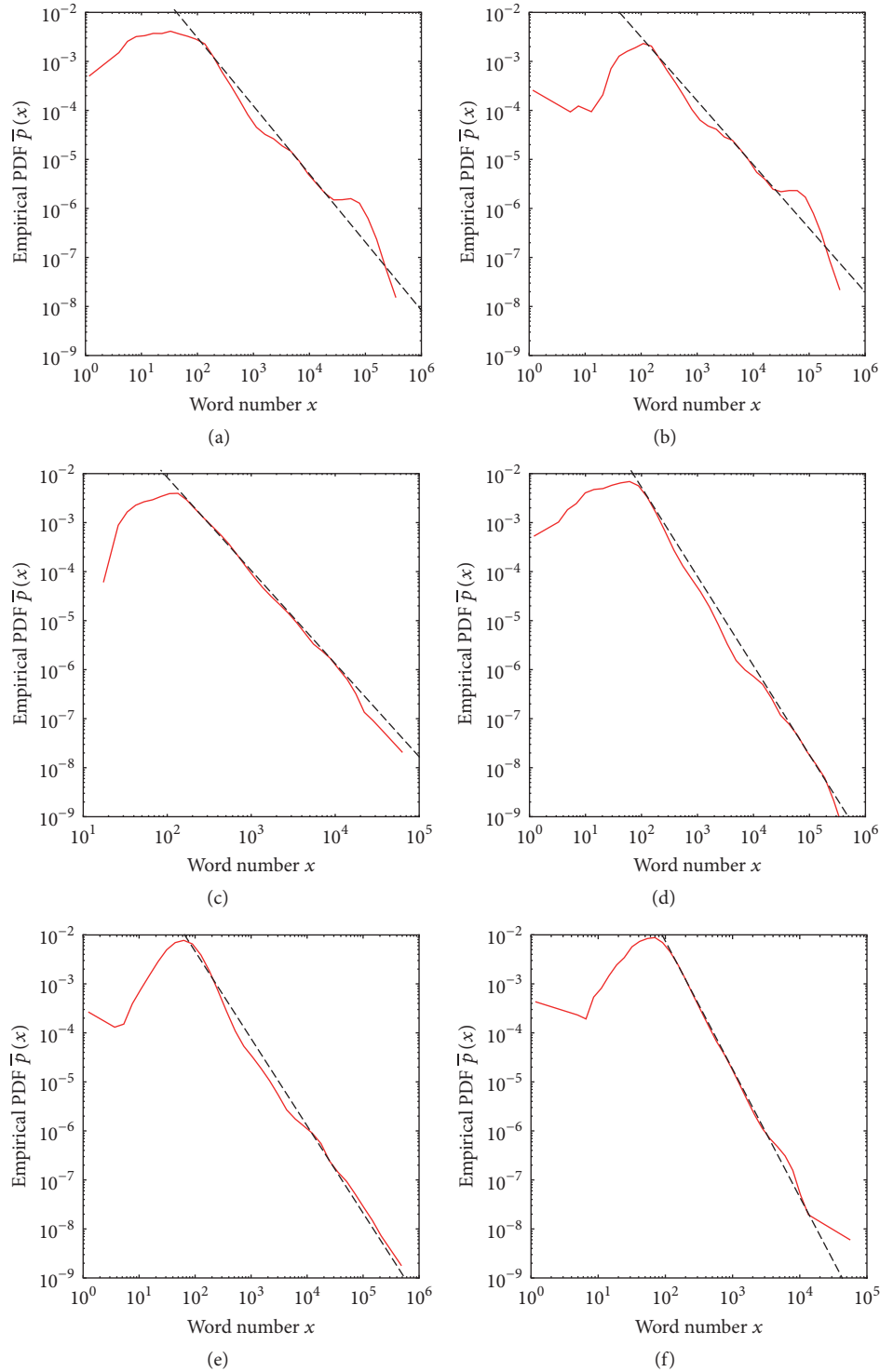


FIGURE 1: The empirical probability density function for the random variable “the number of words in a text” (double-logarithmic scale). (a) English texts, joint sample, (b) English literary works, (c) English poetry, (d) Russian texts, joint sample, (e) Russian literary works, and (f) Russian poetry. Dashed straight lines correspond to power exponents estimated using data collapse.

4. Discussion

Data collapse implies that if the distribution obeys power-law, the transformed distributions possess an interval with horizontal line and coincide inside this interval. For

real-world data, the line inside the interval may not be so straight, but coincidence must occur as Figure 2 shows (one should take into account the fact that the curves are artificially dragged apart a little bit in order to make it possible to

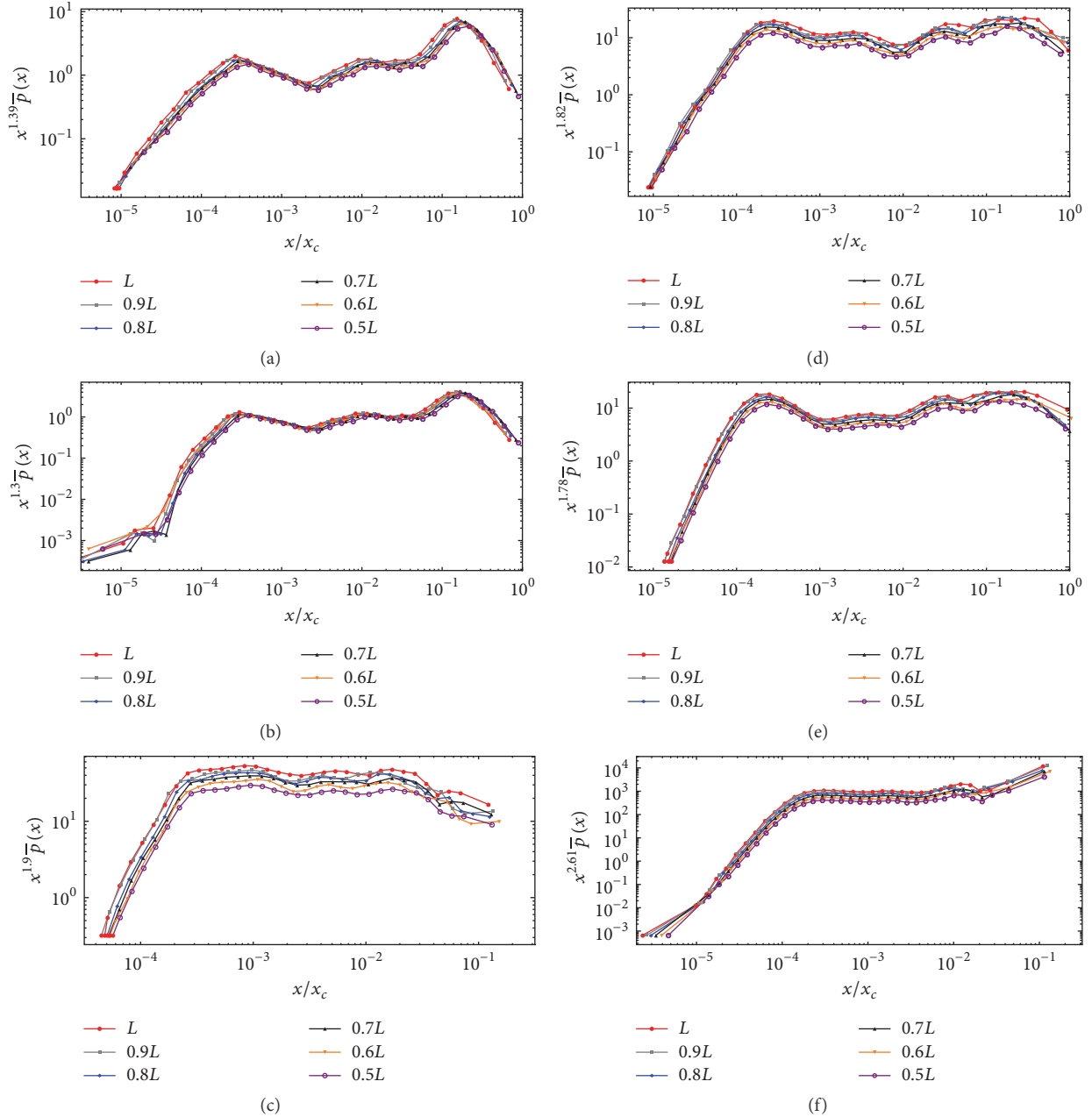


FIGURE 2: Collapsed distributions. (a) English texts, joint sample, (b) English literary works, (c) English poetry, (d) Russian texts, joint sample, (e) Russian literary works, and (f) Russian poetry. Red colour (with discs) stands for the complete vocabulary (of size L), grey colour (with squares) is for the vocabulary of size $0.9L$, blue colour (with diamonds) is for $0.8L$, black colour (with triangles) is for $0.7L$, orange colour (with upturned triangles) is for $0.6L$, and, finally, purple colour (with circles) is for $0.5L$; the curves are artificially dragged apart a little bit in order to make it possible to distinguish them as they are superimposed owing to data collapse. $x^{-\tau}\bar{p}(x)$ versus $\xi = x/x_c$; x_c is a landmark coordinate.

distinguish them as they are superimposed owing to data collapse). Analogously, the results produced by the method that uses KS statistics also count in favour of the hypothesis that distributions are power-law.

We would like to emphasize that we regard this assumption as a plausible hypothesis as before; the results of the previous sections are arguments in its favour, not final results. It seems to us extremely important to ensure global visibility

of this hypothesis. We strongly hope that other papers concerning this hypothesis will occur, with broader data sets and, probably, with more rigorous statistical methods. Fairly good agreement (for this class of distributions) of parameters for separate distributions of literary works created before the 20th century and in the 20th century (Table 2) suggests that a language (at least written language) is a single system diachronically.

TABLE 1: Estimated power-law exponents and lower cutoffs for distributions of the number of words in texts of literary works and social media.

Sample type	<i>English</i>				<i>Russian</i>			
	Estimates based on data collapse		Estimates based on KS criterion		Estimates based on data collapse		Estimates based on KS criterion	
	x_{\min} (nw)	τ (dl)	x_{\min} (nw)	τ (dl)	x_{\min} (nw)	τ (dl)	x_{\min} (nw)	τ (dl)
Literary works	114	1.30	55	1.29	72	1.78	56	1.97
Poetry works	125	1.91	109	1.89	112	2.61	98	2.63
Social media	111	2.65	124	2.63	183	2.10	151	2.15
Joint sample	109	1.39	42	1.34	85	1.82	62	1.95

nw: the number of words; dl: dimensionless.

TABLE 2: Estimated power-law exponents and lower cutoffs for distributions of the number of words in texts for literary works for XIXth (and before it) and for XXth century.

Time limit	<i>English</i>				<i>Russian</i>			
	Estimates based on data collapse		Estimates based on KS criterion		Estimates based on data collapse		Estimates based on KS criterion	
	x_{\min} (nw)	τ (dl)	x_{\min} (nw)	τ (dl)	x_{\min} (nw)	τ (dl)	x_{\min} (nw)	τ (dl)
XIX century and early	118	1.37	75	1.37	91	1.71	53	1.87
XX century	121	1.19	54	1.21	74	1.75	53	2.03

nw: the number of words; dl: dimensionless.

We would like to dwell on a cooccurrence graph as a semantic space in greater detail as opposed to rather popular global syntactic dependency tree and similar structures. In the present paper, a text is considered as a basic unit of a language; thus a sentence is a means to break (rather arbitrary) this semantic “avalanche” down. Global syntactic dependency tree is a great tool to explore this avalanche locally, but as far as it generally fails to reveal cross-sentence semantic dependencies, it does not seem to be the best tool to examine the avalanche as a whole. Therefore cooccurrence graph is a natural choice; an edge belongs to this graph if the words (corresponding to its vertices) belong to the same text. Generally, the underlying structure does not seem to be of principal importance for the problem considered.

In our opinion, this allows one to distinguish results of the present work and those for global syntactic dependency trees and cooccurrence graphs with the employment of random walks [24, 25, 29, 30]. We explore real-world semantic “avalanches” generated by a particular language, while the algorithms using random walks produces artificial “avalanches” with the employment of a graph complying with a natural language. In particular, such motion continues (theoretically) perpetually, whereas the avalanches considered in this article are of finite sizes, explicitly defined by the authors of the respective texts. In order to emphasize fundamental distinction between these approaches, one could draw the following analogy: the approach of the present paper and those associated with random walks are related to study of joint distribution of random variables and of product of

distributions of these variables. Nevertheless, the results of this study are likely to be useful for switcher-random-walks models [30] to estimate realistically switching time.

We would also like to emphasize the difference between the classic Zipf’s laws and the distributions considered in this paper: Zipf studied the laws governing means to represent information, while we attempt to explore laws governing semantic flows and, moreover, semantic flows of a language as a whole.

5. Conclusions

As a result of the above analysis several conclusions may be reached on linguistic systems of English and Russian languages. A language system (given by texts generated in its frameworks) is a self-organized critical system defined on its word cooccurrence graph. Texts of a language are “avalanches” flowing down this graph; the large avalanches correspond to literary works, while the smaller ones are associated with spoken language. A fairly good agreement of parameters for separate distributions of literary works and of social media offers a clearer view of synchronic unity of each linguistic system; on the other hand, an analogous comparison between distributions for literary works of XIXth (and before) and of XXth centuries suggests diachronic unity for the systems. Poetry distributions appear closest to a canonical power-law and therefore poetry may be treated as a kind of supporting column of a language.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors are thankful to Mr. Vladimir Marchenko and to Miss Victoria Ankudinova for the manuscript proofreading and language-editing.

References

- [1] R. M. Rorty, *The Linguistic Turn: Recent Essays in Philosophical Method*, University of Chicago Press, Chicago, Ill, USA, 1967.
- [2] N. A. Chomsky, *The Minimalist Program*, MIT Press, Cambridge, Mass, USA, 1995.
- [3] G. K. Zipf, *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*, Addison-Wesley, Boston, Mass, USA, 1949.
- [4] A. Cohen, R. N. Mantegna, and S. Havlin, "Numerical analysis of word frequencies in artificial and natural language texts," *Fractals*, vol. 5, no. 1, pp. 95–104, 1997.
- [5] R. F. I. Cancho, "When language breaks into pieces A conflict between communication through isolated signals and language," *BioSystems*, vol. 84, no. 3, pp. 242–253, 2006.
- [6] R. Ferrer I Cancho, "Zipf's law from a communicative phase transition," *The European Physical Journal B—Condensed Matter and Complex Systems*, vol. 47, no. 3, pp. 449–457, 2005.
- [7] R. Ferrer I Cancho, R. V. Solé, and R. Köhler, "Patterns in syntactic dependency networks," *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 69, no. 5, Article ID 051915, 2004.
- [8] C. T. Kello and B. C. Beltz, "Scale-free networks in phonological and orthographic word form lexicons," in *Approaches to Phonological Complexity*, De Gruyter Mouton, Berlin, Germany, 2009.
- [9] R. V. Solé, B. Corominas-Murtra, S. Valverde, and L. Steels, "Language networks: their structure, function, and evolution," *Complexity*, vol. 15, no. 6, pp. 20–26, 2010.
- [10] C. T. Kello, G. D. A. Brown, R. Ferrer-i-Cancho et al., "Scaling laws in cognitive sciences," *Trends in Cognitive Sciences*, vol. 14, no. 5, pp. 223–232, 2010.
- [11] P. Medina, E. Goles, R. Zarama, and S. Rica, "Self-organized societies: on the Sakoda model of social interactions," *Complexity*, vol. 2017, Article ID 3548591, 16 pages, 2017.
- [12] U. Eco, *La Struttura Assente*, Tascabili Bompiani, 1980.
- [13] P. Bak, *How Nature Works: The Science of Self-Organized Criticality*, Copernicus Publications, Göttingen, Germany, 1996.
- [14] G. Pruessner, *Self-Organized Criticality*, Cambridge University Press, Cambridge, UK, 2012.
- [15] F. de Saussure, *Course in General Linguistics*, Open Court, 3rd edition, 1986.
- [16] W. J. Ong, *Orality and Literacy: the Technologizing of the Word*, Routledge, Abingdon, UK, 2nd edition, 2002.
- [17] A. Baronchelli, R. Ferrer-i-Cancho, R. Pastor-Satorras, N. Chater, and M. H. Christiansen, "Networks in Cognitive Science," *Trends in Cognitive Sciences*, vol. 17, no. 7, pp. 348–360, 2013.
- [18] S. T. Piantadosi, "Zipf's word frequency law in natural language: a critical review and future directions," *Psychonomic Bulletin & Review*, vol. 21, no. 5, pp. 1112–1130, 2014.
- [19] F. Font-Clos, G. Boleda, and Á. Corral, "A scaling law beyond Zipf's law and its relation to Heaps' law," *New Journal of Physics*, vol. 15, no. 9, 2013.
- [20] R. H. Baayen, *Word Frequency Distributions*, vol. 18 of *Text, Speech and Language Technology*, Kluwer Academic, Dordrecht, Netherlands, 2001.
- [21] R. Ferrer I Cancho, R. V. Solé, and R. Köhler, "Patterns in syntactic dependency networks," *Physical review. E, Statistical, nonlinear, and soft matter physics*, vol. 69, no. 5, p. 051915, 2004.
- [22] R. Ferrer I Cancho, A. Capocci, and G. Caldarelli, "Spectral methods cluster words of the same class in a syntactic dependency network," *International Journal of Bifurcation and Chaos*, vol. 17, no. 7, pp. 2453–2463, 2007.
- [23] A. Barrat, M. Barthélemy, and A. Vespignani, *Dynamical Processes on Complex Networks*, Cambridge University Press, Cambridge, UK, 2008.
- [24] J. D. O. Noh and H. Rieger, "Random walks on complex networks," *Physical Review Letters*, vol. 92, no. 11, p. 118701, 2004.
- [25] P. Allegrini, P. Grigolini, and L. Palatella, "Intermittency and scale-free networks: a dynamical model for human language complexity," *Chaos, Solitons & Fractals*, vol. 20, no. 1, pp. 95–105, 2004.
- [26] D. Offord, *Using Russian: A Guide to Contemporary Usage*, Cambridge University Press, Cambridge, UK, 1996.
- [27] T. Shopen, *Language Typology and Syntactic Description*, vol. 3 of *Language Typology and Syntactic Description*, Cambridge University Press, Cambridge, UK, 2007.
- [28] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM Review*, vol. 51, no. 4, pp. 661–703, 2009.
- [29] J. A. Capitán, J. Borge-Holthoefer, S. Gómez et al., "Local-based semantic navigation on a networked representation of information," *PLoS ONE*, vol. 7, no. 8, Article ID e43694, 2012.
- [30] J. N. Goñi, I. Martincorena, B. Corominas-Murtra, G. Arrondo, S. Ardanza-Trevijano, and P. Villoslada, "Switcher-random-walks: a cognitive-inspired mechanism for network exploration," *International Journal of Bifurcation and Chaos*, vol. 20, no. 3, pp. 913–922, 2010.

Research Article

An Approach for Understanding and Promoting Coal Mine Safety by Exploring Coal Mine Risk Network

Yongliang Deng,^{1,2} Liangliang Song,³ Zhipeng Zhou,⁴ and Ping Liu^{3,5}

¹State Key Laboratory for Geomechanics & Deep Underground Engineering, China University of Mining and Technology, Xuzhou 22116, China

²School of Mechanics and Civil Engineering, China University of Mining and Technology, Xuzhou 221116, China

³School of Civil Engineering, Southeast University, Nanjing 210096, China

⁴College of Economic and Management, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

⁵School of Civil Engineering, Lanzhou University of Technology, Lanzhou 730050, China

Correspondence should be addressed to Liangliang Song; 230129183@seu.edu.cn

Received 26 May 2017; Accepted 22 August 2017; Published 12 October 2017

Academic Editor: Gerard Olivar

Copyright © 2017 Yongliang Deng et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Capturing the interrelations among risks is essential to thoroughly understand and promote coal mining safety. From this standpoint, 105 risks and 135 interrelations among risks had been identified from 126 typical accidents, which were also the foundation of constructing coal mine risk network (CMRN). Based on the complex network theory and *Pajek*, six parameters (i.e., network diameter, network density, average path length, degree, betweenness, and clustering coefficient) were employed to reveal the topological properties of CMRN. As indicated by the results, CMRN possesses scale-free network property because its cumulative degree distribution obeys power-law distribution. This means that CMRN is robust to random hazard and vulnerable to deliberate attack. CMRN is also a small-world network due to its relatively small average path length as well as high clustering coefficient, implying that accident propagation in CMRN is faster than regular network. Furthermore, the effect of risk control is explored. According to the result, it shows that roof collapse, fire, and gas concentration exceeding limit refer to three most valuable targets for risk control among all the risks. This study will help offer recommendations and proposals for making beforehand strategies that can restrain original risks and reduce accidents.

1. Introduction

China is the largest producer and consumer of coal in the world, from which it has derived about 65% of its energy over the past sixty years [1]. In China, more than 90% of fossil energy reserves are coal. That is to say, the energy consumption structure of energy, which relies mainly on coal, cannot be changed within quite a long time. Also, this standpoint can be validated by *China's National Energy Development Strategy Plan (2014–2020)* and *13th Five-Year Plan (2016–2020)*. In 2015, China's coal output was estimated to be 3.747 billion tons, accounting for 47% of the total in the world (The State Administration of Coal Mine Safety, 2015). According to British Petroleum (BP) Statistical Review of World Energy 2016, the countries whose coal production is larger than 40 million tons can be shown in Figure 1.

Coal mining refers to one of the most hazardous industries worldwide [2–4]. Moreover, coal mine enterprises have to encounter various hazards regarding special geological condition [3]. In the process of coal mining, numerous hazards have the potential to trigger accidents frequently, such as rock stresses, harmful gases, humidity, high temperatures, coal and silica dust, and specialized equipment [5]. Worse still, the intensity and frequency of these hazards could result in extremely serious consequences for human health and life [6]. Coal mine accidents will considerably bring about injuries, casualties, and loss of major assets of enterprise. In China, coal mine accident suffers heavy losses every year. According to statistics, approximately 70% of the coal mine casualties worldwide are estimated to occur in China [7]. 6995 coal workers were killed in various accidents in 2002, which is the maximum record in a single year.

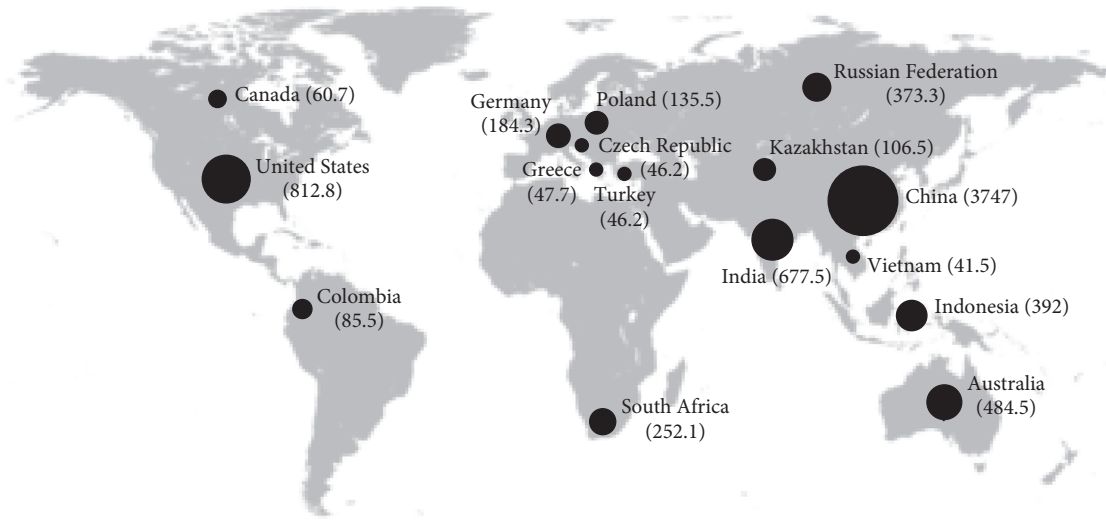


FIGURE 1: The coal production distributed by country in 2015 (one hundred million tons).

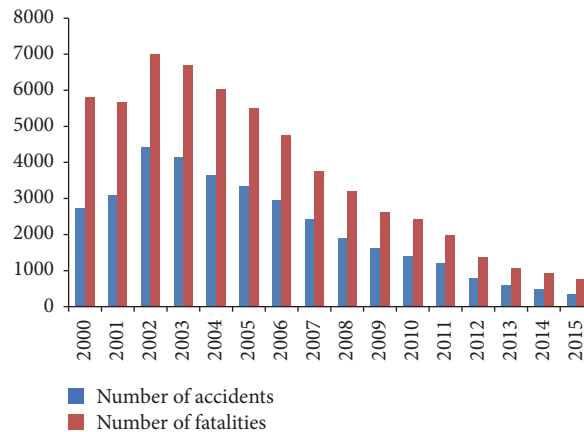


FIGURE 2: China's number of accidents and fatalities in coal mine from 2000 to 2015.

Then, it decreased year by year, as shown in Figure 2 (data source: State Administration of Coal Mine Safety). Although the practical situation seemingly gets better and better, still numerous accidents occurred every year in China. All in all, safety management in coal production is still quite critical and serious due to harsh production conditions as well as complicated production processes.

A more valuable process to improve safety performance is to learn from the failure experiences of previous accidents [8]. Accident analysis is a powerful approach for preventing or eliminating similar hazards, risks, and accidents [9, 10]. Indeed, the existing studies often focus on one type of coal mining accident, or statistical analysis of accident in an area or country, while multiple interrelations among verified accidents are usually neglected. In industrial safety research, it is generally acknowledged that the accident is not caused by a single error or fault, but by the confluence of a sequence of hazard, risk, and accident [11]. Moreover, an occurred accident will possibly incur a sequence of the following accidents [12]. Accident chain exists in most of the coal

mine accidents, which indicates the actual existence of risk network. These interactions among risks form a coal mine risk network (CMRN) which would bring about a big issue for the coal mine safety. Therefore, capturing the complexity of CMRN is both essential and beneficial to improve safety performance in coal mining.

The structure of this paper can be listed as follows. Section 2 presents a literature review of coal mine safety, and Section 3 elaborates the methodology, including an analytical framework, data collection and analysis, and network modeling. In Section 4, *Pajek* is employed to help explore CMRN (including network basic quantities metric and network property) and measure the effect of risk control. In Section 5, the potential contributions, limitations, and risk control methods are discussed. Lastly, the conclusions are drawn in Section 6.

2. Literature Review

Coal mine provides essential energy for supporting high-speed development of Chinese economy and society. Multiple

TABLE 1: Summary of previous study in coal mine.

Theme	References	Objective
Supervision and regulation	[13–18]	Exploring complexity and ineffectiveness of regulation; analyzing rent-seeking mechanism, behavior, policy, and tax; identifying tendency of coal mine accidents and characteristics of human factors.
Risk management	[19–24]	Predicting the expected risk levels by using decomposition technique in time series analysis; analyzing and optimizing the risk management system; using public communication system to monitor unsafe behavior in real time; reducing the effects of coal mining on social and ecological exploitation; constructing potential hazards database in an underground mine; evaluating the reliability of human safety barrier in coal mine emergency evacuation; identifying the risk factors and evaluating the safety control capability.
Risk evaluation	[25–31]	Assessing the roof fall risk during retreat mining in room and pillar coal mines; evaluating explosion risk in underground coal mines; developing a comprehensive model for coal mine safety; using fuzzy set theory to assess the risk of mining equipment failure; assessing pot-hole subsidence risk in coal mine; using risk performance indicators to analyze coal mine accidents.
Monitoring and controlling technologies	[32–36]	Employing internet of things (IOT) and cloud computing (CC) to monitor mine safety based on prealarm system; using wireless sensor network (WSN) to monitor the temperature, humidity, gas, and status of smoke in underground mine; establishing a Web of Things-based remote monitoring system for coal mine safety; employing cable monitoring system (CMS) and the WSN to build an integrated environment monitoring system for underground coal mine; using iris identification and radio frequency identification (RFID) technique to improve safety management system.

studies have been carried out by worldwide researchers to improve the safety performance. The research topics mainly focus on supervision and regulation, risk management, evaluation, monitoring, and controlling technologies, which is shown in Table 1.

Supervision and regulation refer to two crucial influence factors in the coal mining. Before 2000, ineffective implementation of laws and regulations increases the difficulty for Chinese government to inspect actual situation of coal mine safety [13]. To promote coal mine safety, a variety of effective countermeasures, such as enhancing safety legislation and establishing independent coal mine safety monitoring system, were executed. These improvements in regulatory regime make a great contribution [14]. However, the interrelations between coal mine enterprises and supervision departments are complex and subtle. Rent-seeking exists widely in China's coal mine supervision, which is a huge obstacle to the further development of coal mining industry. The existing researches on rent-seeking mainly focus on rent-seeking behavior, policy, and tax [15–17]. In the rent-seeking scenario, Chen et al. [18] indicated that each level of the department had an intensity threshold above which coal mine accidents occurred.

The effective risk management is the fundamental guarantee of coal mine safety production based on various theories and methods. Sari et al. [19] developed a stochastic model to predict the number of accidents according to the randomness in the occurrence of accidents. Qing-gui et al. [20] constructed a system to supervise unsafe behavior, release early warning information, and improve controlling measures in coal mine. Based on case studies, Kowalska [21] identified and assessed the risk sources. As suggested

by the results, it is necessary to undertake anticipatory activities aiming at reducing environmental and social risks during the colliery liquidation. Badri et al. [22] studied risk management in mining projects based on analytic hierarchy process method, and the results show the importance of considering occupational health and safety (OHS) in the process of coal mining. Wang et al. [23] put forward an analytical framework to analyze human error risk in the emergency evacuation from three perspectives, including organization level, group level, and individual level. Besides, Liu and Li [24] constructed a back propagation (BP) neural network to explore influence factors in coal mine safety.

The evaluation of hazards and risks has attracted much attention of multiple practitioners and researchers due to their serious consequences. These hazards and risks could be divided into three types, including “natural, technical, and human.” Ghasemi et al. [25] developed a risk evaluation model and various possible risks are evaluated in Iran Tabas central mine. Pejic et al. [26] proposed a risk assessment tool to determine the risk of explosion of any work processes or activities in the underground coal mine. Also, the methodology can decide whether the proposal investments are well-justified or not for improving safety. Bahri Najafi et al. [27] proposed an artificial neural network model to predict the out-of-seam dilution. Based on uncertain random variables, Chen et al. [28] developed a practical evaluation model for coal mine safety based on uncertain random variables. According to fuzzy set theory, Petrović et al. [29] presented a risk evaluation model to evaluate the failures of electromechanical equipment. Lokhande et al. [30] came up with a risk evaluation approach based on the identified critical parameters, including depth to height of extraction

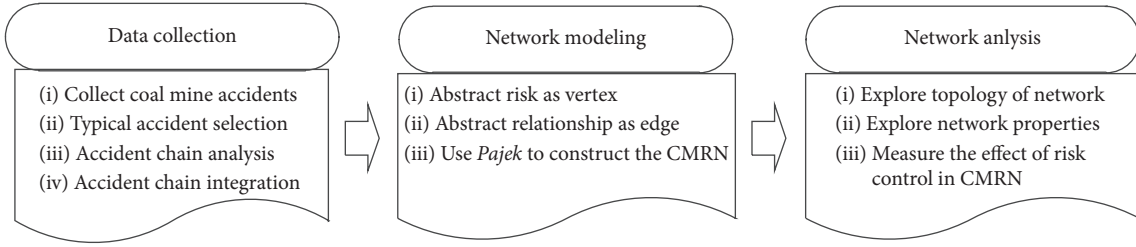


FIGURE 3: Analytical framework.

TABLE 2: Two examples of the stored accidents.

Number	Time	City	Type	Description	Death	Loss
28	2003.10.21	Wuhai	Coal dust explosion	Three blasters violated job regulations and implemented blasting without any safety precautions. Unfortunately, the blasting gave rise to naked light, and then the gas was lighted and began burning. As a result, the coal dust explosion happened.	6	120 thousand dollars
92	2004.2.23	Jixi	Gas explosion	Due to inadequate ventilation, the gas concentration exceeded the threshold. Meanwhile, a miner optionally disassembled his lamp, which triggered electric spark, and then the spark caused gas explosion.	37	370 thousand dollars

ratio, rock to soil ratio, brittleness index of rock, and rock density. Spada and Burgherr [31] analyzed the accident data in energy-related severe accidents database and suggested a nonsignificant decreasing tendency for Turkey as well as a significant one for USA.

Some new technologies, which are effective and powerful tools for improving safety performance, have been applied in coal mine. Sun et al. [32] accomplished a monitoring and prealarm system based on cloud computing (CC) and Internet of things (IOT). What is more, Dange and Patil [33] designed a wireless sensor network (WSN) based on MSP430 controller for monitoring smoke, gas, temperature, and humidity in coal mine. Based on wireless sensor network and controller area network (CAN), Bo et al. [34] proposed a remote monitoring system, which was tested in different remote monitoring scenarios. Zhang et al. [35] proposed an integrated environment monitoring system that takes full advantage of cable monitoring system (CMS) in combination with wireless sensor network (WSN). Xu et al. [36] put forward an improved safety management system based on several modern identification and communication techniques, including iris identification, radio frequency identification (RFID), computer network, and database technique.

3. Methodology

3.1. Analytical Framework. An analytical framework is proposed to conduct the in-depth analysis of coal mine accident, as presented in Figure 3. It is a step-by-step procedure consisting of three main modules. At first, the coal mine accidents are collected from literature and media, such as the website of State Administration of Coal Mine Safety. Then, typical accidents are selected as the data to analyze accident chains. After that, the accident chains will be integrated as

a global network. In the second stage, the risk is abstracted as vertex, and meanwhile, the interrelation is abstracted as edge. Also, the software *Pajek* is employed to establish the coal mine risk network (CMRN). In the third stage, the topology of CMRN is analyzed and network properties are identified according to the network theory. Then, the effect of risk control in CMRN is calculated. According to the research result, the discussions and suggestions are provided to promote safety management in coal mine production.

3.2. Data Collection and Analysis. The data of historical coal mining accidents is used for risk analysis. There are several ways to collect accident cases, such as government, enterprise, literature, and media. In this study, the accidents are collected from literature and media. A coal mine accident database (CMAD), which records the detailed information of accident (including time, position, type, process, death, and losses), is established based on Microsoft Access 2010. Although hundreds of accidents have occurred in China over the past few years, the information of many accidents, especially the process of accident, is unclear. In the end, 176 accidents with exhaustive information are collected. Among these detailed accidents, some accident chains are unobvious, while some happen suddenly and unexpectedly without accident chain. These accidents are not considered in this research. Besides, since some accidents are exactly similar to the rest of the typical accidents, thus there is no need to analyze the repeating accidents. In the end, 126 typical accidents, including all types of coal mining accidents, are recorded in CMAD, and they are selected to conduct accident chain analysis for establishing the risk network model. Two examples of stored accidents can be illustrated in Table 2.

Although these accidents are selective, almost all kinds of accidents have been included. Also, there are no biases in the

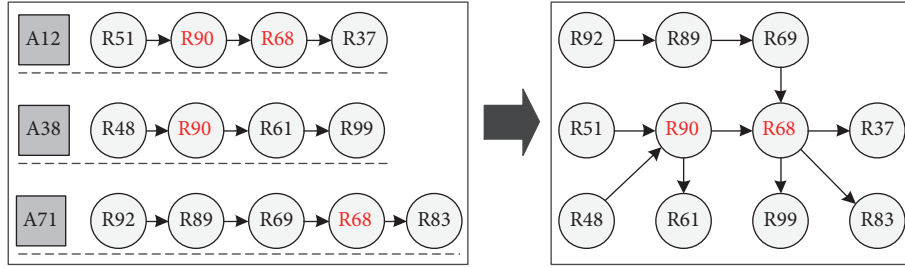


FIGURE 4: The formative process of CMRN.

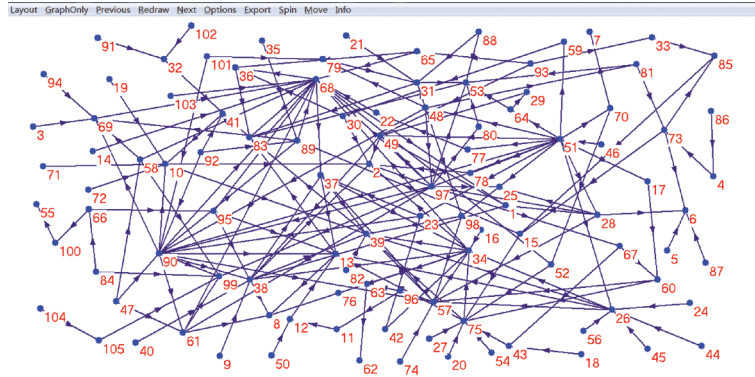


FIGURE 5: The coal mine risk network model in Pajek.

selection process. From the perspective of person, machine, environment, management, and technology, the accident chains in these accidents are identified and expatiated in Table 3. Most of the accidents have one accident chain, while some have two, such as accidents 41 and 75. As a result, a total of 135 accident chains are obtained from 126 cases.

3.3. Network Modeling. Multiple risks simultaneously appear in different accidents, indicating that the risk is correlated with others. It is essential to identify the risks and interrelations among them so as to establish CMRN. Through statistics, a total of 105 risks and 194 interrelations are obtained from 135 accident chains. Moreover, the vertex number and its type are expatiated in Table 4. After this study abstracts risk as vertex and interrelation as edge, different risks can be connected by these common vertexes into a global network. For a better explanation, accidents 12, 38, and 71 are taken as an example to illustrate the process of network modeling, as depicted in Figure 4. From the risks identification in accidents 12, 38, and 71, it can be seen that there are two same vertexes shown in red color, including R90 and R68. Through this method, the network can be established based on these common risks. Furthermore, software *Pajek* is employed to establish coal mine risk network (CMRN), as shown in Figure 5.

4. Results

4.1. Network Basic Quantities Metric. With the continuous development of complex network theory, the statistical indexes of network structure have obtained a lot of

achievements, which are also the basis of statistical description of various topological characteristics. Compared with visual section, the calculation is much more precise and concise in exploring network [37]. This study uses several typical indexes to explore the properties of CMRN, including network diameter, network density, average path length, degree, betweenness, and clustering coefficient. These topological indexes are calculated by *Pajek*.

4.1.1. Network Diameter. The network diameter is defined as the maximum path length in the network, which can reflect the size of a network. The network diameter in CMRN is 7, which is from poor maintenance (vertex 64) to water leaking (vertex 99). This path is as follows: poor maintenance (vertex 64) causes electrified device failure (vertex 29), electrified device failure (vertex 29) triggers inadequate ventilation (vertex 49), inadequate ventilation (vertex 49) makes gas concentration exceed limit (vertex 38), gas concentration exceeding limit (vertex 38) incurs gas burning (vertex 37), gas burning (vertex 37) sparks off fire (vertex 34), fire (vertex 34) induces roof collapse (vertex 68), roof collapse (vertex 68) leads to penetration into goaf (vertex 61), and penetration into goaf (vertex 61) brings about water leaking (vertex 99). Although these risks may not occur simultaneity in a single accident, it can deeply reflect the process of risk spread. The spread rule of risk is conducive to developing prevention and control strategies for the risk control.

4.1.2. Network Density. Network density is used to describe the degree of affinity between the vertexes in a network from

TABLE 3: Accident chain analysis.

Number	Accident chains
1	Smoking → naked light → fire → suffocation
2	Unreasonable blasting → gas concentration exceeding limit → gas burning → fire
3	Air blower failure → inadequate ventilation → gas concentration exceeding limit → gas burning → fire → gas explosion
4	Electric leakage → gas burning → gas explosion
5	Coal and gas outburst → suffocation
6	Electrical failures → air blower failure → inadequate ventilation → Gas concentration exceeding limit → fire
7	Violation operation → inadequate ventilation → gas concentration exceeding limit → gas explosion
8	Inadequate ventilation → gas concentration exceeding limit → suffocation
9	Violation operation → suffocation
10	Management negligence → violation weld → gas explosion
11	Management negligence → inadequate ventilation → suffocation
12	Management negligence → unreasonable blasting → roof collapse → gas burning
13	Mechanical friction → spark → gas explosion
14	Air blower failure → inadequate ventilation → gas concentration exceeding limit → gas explosion
15	Electric spark → gas burning → fire → gas explosion
16	Management negligence → violation operation → suffocation
17	Roof collapse → mechanical friction → spark → gas explosion
18	Inadequate ventilation → gas concentration exceeding limit → gas explosion
19	Unreasonable technique scheme → geostress concentration → coal and gas outburst → suffocation
20	Imperfect regulation → management negligence → unreasonable blasting → coal and gas outburst → suffocation
21	Management negligence → poor maintenance → electrical device failure → inadequate ventilation → gas concentration exceeding limit → gas explosion
22	Unreasonable blasting → coal dust explosion → carbon monoxide poisoning
23	Imperfect regulation → management negligence → electric spark → coal dust explosion
24	Management negligence → ruptured steel rope → mechanical friction → spark → gas explosion
25	Ruptured steel rope → mechanical friction → spark → coal dust explosion
26	Unreasonable blasting → roof collapse → ventilation failure → coal dust explosion → struck-by
27	Broken steel rope → sliding train → collision → spark → coal dust explosion
28	Violation operation → unreasonable blasting → naked light → gas burning → coal dust explosion
29	Electric spark → naked light → gas burning → fire → carbon monoxide poisoning
30	Electrical failures → naked light → fire → roof collapse
31	Cable short circuit → electric spark → fire → carbon monoxide poisoning
32	Electrical failures → cable short circuit → electric spark → fire → carbon monoxide poisoning
33	Violation operation → pressure fan failure → over-temperature → spark → fire → carbon monoxide poisoning
34	Inadequate training → violation weld → naked light → fire → carbon monoxide poisoning
35	Management negligence → electric spark → naked light → fire → roof collapse
36	Management negligence → conveyor failure → over-temperature → naked light → fire → suffocation
37	Management negligence → electrical failures → air blower failure → gas concentration exceeding limit → carbon monoxide poisoning
38	Inadequate training → unreasonable blasting → penetration into goaf → water leaking
39	Violation operation → penetration into goaf → water leaking
40	Management negligence → unreasonable blasting → penetration into goaf → water leaking
41	Unreasonable blasting → gas concentration exceeding limit
42	Unreasonable blasting → naked light → gas burning → gas explosion
43	Inadequate geological prospecting → unreasonable blasting → penetration into goaf → water leaking
43	Inadequate geological prospecting → penetration into goaf → water leaking

TABLE 3: Continued.

Number	Accident chains
44	Penetration into goaf → water leaking
45	Management negligence → violation operation → water leaking
46	Management negligence → stagnant water → suffocation
47	Management negligence → violation operation → penetration into goaf → water leaking
48	Inadequate training → standing on conveyor belt → fall → mechanical injury
49	Management negligence → ruptured steel rope → falling of cage
50	Broken steel rope → sliding train → train derailment → collision
51	Inadequate training → broken steel rope → mechanical injury
52	Unscientific design → faulty track → train derailment → collision
53	Train overload → brake failure → sliding train → train derailment → collision
54	Management negligence → poor maintenance → mechanical injury
55	Violation operation → electrical failures → electric shock
56	Management negligence → violation operation → electric shock
57	Cable short circuit → electric shock
58	Management negligence → inadequate training → fall → mechanical injury
59	Stray current → gas explosion
60	Management negligence → electric leakage → unreasonable blasting
61	Unscientific design → poor tunnel support → roof collapse
62	Float coal → tunnel support failure → roof collapse
63	Train derailment → collision → tunnel support failure → roof collapse
64	Optional withdrawal of pillar → roof collapse → struck-by
65	Inadequate geological prospecting → neglect of geostress concentration → roof collapse
66	Neglect of geostress concentration → roof separation → roof collapse
67	Poor tunnel support → flying rock → struck-by
68	Unreasonable blasting → tunnel support failure → roof collapse
69	Unscientific design → poor tunnel support → roof collapse → flying rock
70	Violation operation → roof collapse
71	Unreasonable technique scheme → tunnel support failure → roof separation → roof collapse → struck-by
72	Unreasonable technique scheme → unreasonable blasting → roof separation → roof collapse
73	Poor tunnel support → roof collapse → flying rock
74	Weakness of safe consciousness → standing on conveyor belt → fall → mechanical injury
75	Roof collapse → penetration into goaf → carbon monoxide poisoning
76	Air blower failure → inadequate ventilation → gas concentration exceeding limit Electrical failures → cable short circuit → electric spark → fire
77	Inadequate geological prospecting → neglect of geostress concentration → water leaking
78	Weakness of safe consciousness → unreasonable blasting → water leaking
79	Management negligence → spontaneous combustion of dynamite → dynamite explosion → roof collapse
80	Transformer overload → cable short circuit → electric spark → fire
81	Sudden torrential rain storm → water leaking
82	Spontaneous combustion of coal seam → fire
83	Unreasonable blasting → naked light → fire
84	Violation operation → electric shock
85	Inadequate training → violation operation → mechanical injury
86	Unreasonable blasting → coal and gas outburst → gas explosion
87	Drilling blasting hole → spark → gas explosion

TABLE 3: Continued.

Number	Accident chains
88	Violation blasting → naked light → gas explosion → struck-by
89	Miner's lamp failure → electric spark → gas explosion
90	Unscientific design → inadequate ventilation → gas concentration exceeding limit Cable insulation failure → cable short circuit → electric spark → gas explosion
91	Smoking → naked light → gas explosion
92	Inadequate ventilation → gas concentration exceeding limit Illegal disassembly of miner's miner's lamp → electric spark → gas explosion
93	Defective geological condition → coal and gas outburst Electric locomotive failure → electric spark → gas explosion
94	Gas monitoring system failure → gas concentration exceeding limit → gas explosion
95	Lack of dedusting device → coal dust concentration exceeding limit → coal dust explosion
96	Violation blasting → collapse of coal bunker → coal dust concentration exceeding limit → coal dust explosion
97	Violation weld → conveyor belt burning → fire → suffocation
98	Sudden torrential rain storm → power cut → water pump failure → mine flooding
99	Power cut → ventilation failure → gas concentration exceeding limit Violation blasting → naked light → gas explosion
100	Violation blasting → poisonous gas leakage → poisoning
101	Severe vibration in coal cutting → coal and gas outburst
102	Explosion of electric switch → spark → gas explosion
103	Roof collapse → air blower failure → inadequate ventilation → gas concentration exceeding limit Cable short circuit → electric spark → fire
104	Metal crash → spark → gas explosion
105	Illegal restart → electric spark → gas explosion
106	Optional close of ventilation → gas concentration exceeding limit Illegal disassembly of miner's miner's lamp → electric spark → gas explosion
107	Severe vibration in anchor construction → coal and gas outburst
108	Broken steel rope → sliding train → cable short circuit → electric spark → coal dust explosion
109	Mechanical friction → spark → coal dust explosion
110	Management negligence → electric leakage → ignition of cable → fire → poisonous gas leakage → poisoning
111	Conveyor over-temperature → ignition of engine oil → spark → fire
112	Violation operation → water leaking
113	Wrong geologic survey → wrong holing-through → water leaking
114	Delay of support → roof separation → roof collapse
115	Pressure fan failure → ignition of engine oil → spark → fire
116	Unstable pillar → roof separation → roof collapse
117	Winch brake failure → falling object → struck-by → mechanical injury
118	Trip → fall → struck-by
119	Trip → mechanical injury
120	Collapse of support structure → roof collapse → struck-by
121	Drinking → fall → mechanical injury
122	Violation operation → steel rope bouncing → mechanical injury
123	Management negligence → no warning sign Flying rock → struck-by
124	Management negligence → no warning sign → entering danger zone → suffocation
125	Entering danger zone → flying rock → struck-by
126	Unreasonable dismantling of elevator → falling object → struck-by

TABLE 4: Risk number and type.

Number	Risk	Attribute
1	Air blower failure	Machine
2	Asphyxiation	Environment
3	Delay of support	Management
4	Brake failure	Machine
5	Cable insulation failure	Machine
6	Cable short circuit	Machine
7	Falling of cage	Machine
8	Carbon monoxide poisoning	Environment
9	Optional close of ventilation	Management
10	Coal and gas outburst	Environment
11	Collapse of coal bunker	Environment
12	Coal dust concentration exceeding limit	Environment
13	Coal dust explosion	Environment
14	Collapse of support structure	Technology
15	Collision	Machine
16	Conveyor belt burning	Machine
17	Conveyor failure	Machine
18	Conveyor over-temperature	Machine
19	Defective geological condition	Environment
20	Drilling blasting hole	Technology
21	Drinking	Person
22	Dynamite explosion	Management
23	Electric leakage	Machine
24	Electric locomotive failure	Machine
25	Electric shock	Machine
26	Electric spark	Machine
27	Explosion of electric switch	Machine
28	Electrical failure	Machine
29	Electrified device failure	Machine
30	Entering danger zone	Person
31	Fall	Person
32	Falling object	Environment
33	Faulty track	Machine
34	Fire	Environment
35	Float coal	Environment
36	Flying rock	Environment
37	Gas burning	Environment
38	Gas concentration exceeding limit	Environment
39	Gas explosion	Environment
40	Gas monitoring system failure	Machine
41	Geostress concentration	Environment
42	Ignition of cable	Machine
43	Ignition of engine oil	Machine
44	Illegal disassembly of miner's miner's lamp	Person
45	Illegal restart	Person
46	Imperfect regulation	Management

TABLE 4: Continued.

Number	Risk	Attribute
47	Inadequate geological prospecting	Management
48	Inadequate training	Management
49	Inadequate ventilation	Management
50	Lack of dedusting device	Management
51	Management negligence	Management
52	Mechanical friction	Machine
53	Mechanical injury	Machine
54	Metal crash	Machine
55	Mine flooding	Environment
56	Miner's lamp failure	Machine
57	Naked light	Machine
58	Neglect of geostress concentration	Management
59	No warning sign	Management
60	Over-temperature	Environment
61	Penetration into goaf	Technology
62	Poisoning	Environment
63	Poisonous gas leakage	Environment
64	Poor maintenance	Management
65	Poor tunnel support	Technology
66	Power cut	Machine
67	Pressure fan failure	Machine
68	Roof collapse	Environment
69	Roof separation	Environment
70	Ruptured steel rope	Machine
71	Severe vibration in anchor construction	Technology
72	Severe vibration in coal cutting	Technology
73	Sliding train	Machine
74	Smoking	Person
75	Spark	Machine
76	Spontaneous combustion of coal seam	Environment
77	Spontaneous combustion of Dynamite	Management
78	Stagnant water	Environment
79	Standing on conveyor belt	Person
80	Steel rope bouncing	Machine
81	Broken steel rope	Machine
82	Stray current	Machine
83	Struck-by	Person
84	Sudden torrential rain storm	Environment
85	Train derailment	Machine
86	Train overload	Machine
87	Transformer overload	Machine
88	Trip	Person
89	Tunnel support failure	Technology
90	Unreasonable blasting	Technology
91	Unreasonable dismantling of elevator	Technology
92	Unreasonable technique scheme	Technology

TABLE 4: Continued.

Number	Risk	Attribute
93	Unscientific design	Technology
94	Unstable pillar	Technology
95	Ventilation failure	Machine
96	Violation blasting	Management
97	Violation operation	Management
98	Violation weld	Management
99	Water leaking	Environment
100	Water pump failure	Machine
101	Weakness of safe consciousness	Person
102	Winch brake failure	Machine
103	Optional withdrawal of pillar	Management
104	Wrong geologic survey	Technology
105	Wrong holing-through	Technology

an overall perspective. It specifically refers to the proportion of actual edges to potential edges in a network. Consisting of 105 vertexes, the maximum number of edges in CMRN should be $105 * 104 = 10920$. Since the actual edges in CMRN is 194, thus the network density of CMRN is $194/10920 = 0.178$. In general, the more the vertexes, the smaller the network density. Low density means that CMRN is a relatively sparse network. Moreover, the vertex in CMRN is less connected with all others. That is to say, the degree of a vertex in CMRN directly affected by others is relatively low.

4.1.3. Average Path Length. The transmission efficiency of information or energy is significantly correlated with the average path length. A shorter average path length means higher efficiency. The average path length can be defined as the average number of steps between all possible pairs of vertexes in a network. The value of the average path length in CMRN is 3.0841, indicating that a risk can transmit to another only in three steps on average. For example, cable short circuit (vertex 6) and carbon monoxide poisoning (vertex 8) refer to two correlative risks, which can be connected by electric spark (vertex 26) and fire (vertex 34) in three steps, as shown in accident 31 in Table 3.

4.1.4. Degree. The degree of a vertex is defined as the number of edges connected to the vertex. In a directed network, the degree can be either in-degree (number of incoming edges) or out-degree (number of outgoing edges), with the total degree being the sum of the two. Since there are 105 vertexes in CMRN, it is impossible to show all the vertex degree in a radar graph. Consequently, 30 vertexes with the highest degree are selected as the example to display vertex degree. The values of the in-degree, out-degree, and total degree of these 30 vertexes are presented in Figure 6. Roof collapse (vertex 68) has the highest degree of 17, with an in-degree 10 and out-degree 7. This indicates that the roof collapse is in a relatively central position and plays a critical role in the accident chain. Its in-degree is also the highest in the network, implying that it refers to the biggest “risk recipient”

in CMRN and many risks such as poor tunnel support can lead to roof collapse. Multiple paths make it difficult to control for roof collapse, compared to other vertexes with low in-degree. The second is unreasonable blasting (vertex 90) and the third is the management negligence (vertex 51). The in-degree and out-degree of unreasonable blasting are 7 and 9, respectively. It means that 7 risks could give rise to unreasonable blasting, and meanwhile, unreasonable blasting might cause 9 risks in production. Additionally, the management negligence (vertex 51) has the highest out-degree, demonstrating that management negligence is the most serious risk source. If there is something wrong in safety management, many risks might be triggered at any time, such as gas concentration exceeding limit. Controlling these key vertexes can positively influence the safety of coal mine, which is also referential in resource distribution under the condition of limited security resource. Besides, it would greatly help disrupt the connectivity among risks to prevent risks from spreading and propagating in CMRN.

4.1.5. Betweenness. Betweenness is used to describe the extent to which a vertex plays an intermediary role in the interaction between all possible pairs of vertexes in a network [38]. Two types of betweenness, vertex betweenness and edge betweenness, are used extensively in the network analysis [39, 40]. According to the research object, only vertex betweenness is utilized in this study. High betweenness indicates greater importance in the whole network. The vertex betweenness in the CMRN ranges from 0 to 0.059852, as shown in Table 5. Only 47 vertexes are invisible because their vertex betweenness is zero, which indicates that they do not play the role of intermediary among interactions between other vertexes. The roof collapse (vertex 68) has the highest value of vertex betweenness, meaning that the maximum number of the shortest paths passes through roof collapse (vertex 68). It is a key link in the process of risk spread. The stagnant water (vertex 78) has the lowest value of vertex betweenness, meaning that the minimum number of the shortest paths passes through stagnant water (vertex 78). It is not a key link in the process of risk spread. According to the value of betweenness, the impact of roof collapse (vertex 68) is much larger than stagnant water (vertex 78) in the process of risk spread. Furthermore, fire (vertex 34) and spark (vertex 75) are 0.048486 and 0.020668, respectively. The cumulative vertex betweenness of the five highest vertex betweenness is equal to 0.542701, which indicates that about 55% shortest paths pass through these five vertexes. These vertexes should be focused in the safety management. It seems that effectively controlling these few key vertexes can slow down the risk diffusion and decrease the chain reaction in CMRN.

4.1.6. Clustering Coefficient. The clustering coefficient is used to describe which vertexes in a network tend to cluster together from a local perspective [41]. The clustering coefficient of a vertex is defined as the probability of two randomly selected neighbors of the vertex being connected. It can be found that 33 vertexes get the missing value of 999999998 because the degrees of these vertexes are equal to 1, and 34 vertexes have the value of 0. The clustering coefficients of

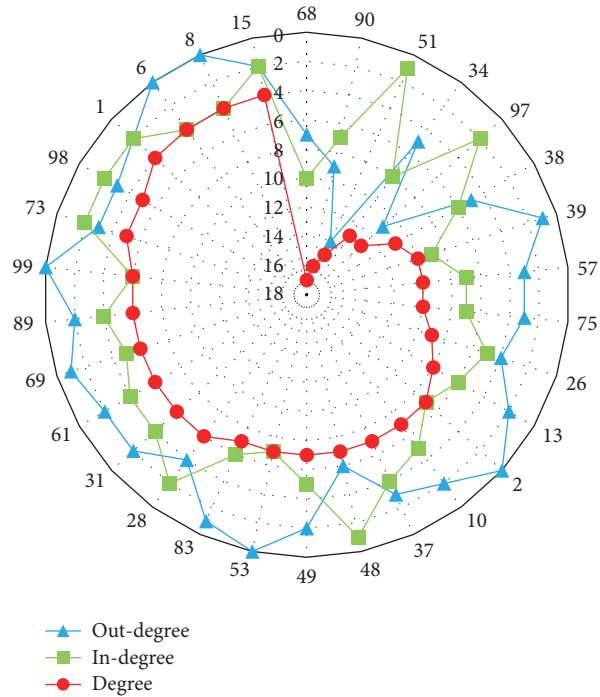


FIGURE 6: In-degree, out-degree, and total degree values.

other 38 vertexes are presented from high to low in Table 6. The clustering coefficient of vertex in CMRN ranges from 0 to 0.5. The vertexes with the highest clustering coefficient are vertex 25 and vertex 80. The network clustering coefficient can be defined as the average value of all vertexes in the network, and it is 0.0623 in CMRN which is larger than a random network with the same network scale. The large clustering coefficient denotes that CMRN has a high degree of cliquishness.

4.2. Network Property. With the development of network theory, it can be found that small-world property and scale-free property are the most obvious distinction between real network and random network. To obtain greater insight into the nature of CMRN, this section explores these two properties.

4.2.1. Small-World Property. A small-world network is a special kind of graph, in which most vertexes can be reached from every other vertex by a short path. In general, small-world network is associated with the possession of relatively high value of clustering coefficient and small average path length [42, 43]. For comparison, three random networks with 105 vertexes and 194 edges are created by *Pajek*, which are the same scale as CMRN. The clustering coefficient and average path length of CMRN and random networks are presented in Table 7. Obviously, CMRN is a relatively small-world network according to its clustering coefficient and average path length, indicating that the risk propagation in CMRN is much faster than a random network. To avoid a worse consequence under the condition of an occurred accident, controlling the catenation among accidents is of great significance.

4.2.2. Scale-Free Property. A scale-free network is a network whose degree distribution satisfies power-law decay. In such network, numerous vertexes are poorly connected and relatively few vertexes are linked to many other vertexes [44]. Due to rare vertexes with high degree, analyzing statistic data in the tail of the degree distribution is meaningless. The degree distribution $P(k)$ is defined as the proportion of vertexes with degree k , while the cumulative $P(k)$ is defined as the proportion of vertexes equaling to or greater than k [45]. In practice, the cumulative $P(k)$ is preferred in statistical analysis using double logarithmic coordinate system, with the purpose of reducing statistical errors caused by finite network size [46]. The cumulative $P(k)$ of CMRN is depicted in Figure 7 with approximate fit $P(k) = 2.1217 \times k^{-1.545}$, which basically follows the power-law. This indicates that the CMRN has scale-free property according to complex network theory. The property means that CMRN is robust to random risks to some extent. The vertex with degree equaling to or less than 4 accounts for 75%, and the influence of these vertexes on the network is relatively small. However, CMRN is vulnerable to simultaneous attacks aiming at vertexes with high degree. In other words, only targeted actions can greatly prevent the cascading effects in CMRN.

4.3. Measuring the Effect of Risk Control. The analysis on effect of risk control is conducive to providing recommendations and proposal for safety management in coal mine. To measure the effect of risk control, an assumption is made. Namely, a risk would be supposed not to occur if it is completely controlled in coal mine production. Furthermore, if a risk will not happen, it can be deleted from CMRN. Then, the effect of risk control can be measured by network global

TABLE 5: Betweenness in CMRN.

Vertex	Betweenness
68	0.059852
34	0.048486
75	0.020668
38	0.016303
15	0.012481
89	0.010105
90	0.009873
57	0.009704
26	0.008742
1	0.008044
73	0.007534
37	0.007143
83	0.007001
61	0.006588
95	0.006566
63	0.006068
52	0.005903
39	0.005813
51	0.005134
13	0.005128
49	0.004844
69	0.003734
60	0.003423
97	0.002563
43	0.002558
85	0.002474
4	0.002427
66	0.001960
10	0.001875
22	0.001867
29	0.001774
48	0.001147
65	0.000996
81	0.000937
28	0.000850
12	0.000794
16	0.000770
98	0.000745
31	0.000742
67	0.000660
36	0.000560
58	0.000490
70	0.000467
30	0.000436
33	0.000420
23	0.000373
32	0.000373
17	0.000280
64	0.000280
59	0.000249
79	0.000218
77	0.000187
100	0.000187
42	0.000179
11	0.000140
41	0.000101

TABLE 5: Continued.

Vertex	Betweenness
105	0.000093
78	0.000062
:	0

TABLE 6: Clustering coefficient in CMRN.

Vertex	Clustering coefficient
25	0.5000
80	0.5000
88	0.5000
61	0.2000
36	0.1667
47	0.1667
65	0.1667
79	0.1667
85	0.1667
92	0.1667
37	0.1429
69	0.1333
89	0.1333
48	0.1190
31	0.1000
39	0.1000
98	0.1000
99	0.1000
49	0.0952
1	0.0833
8	0.0833
15	0.0833
23	0.0833
58	0.0833
97	0.0833
2	0.0714
53	0.0714
34	0.0705
28	0.0667
57	0.0667
38	0.0636
90	0.0583
73	0.0500
83	0.0476
26	0.0417
68	0.0368
51	0.0333
75	0.0111
:	0

efficiency. Although many definitions on network global efficiency are currently created and studied, they all have different limitations. The generally accepted measure method is average reciprocal shortest path lengths of networks, in

TABLE 7: The comparison between CMRN and random networks.

Network model	Clustering coefficient	Average path length
CMRN	0.0623	3.0841
Random network 1	0.0156	5.6134
Random network 2	0.0189	5.8130
Random network 3	0.0152	5.4532

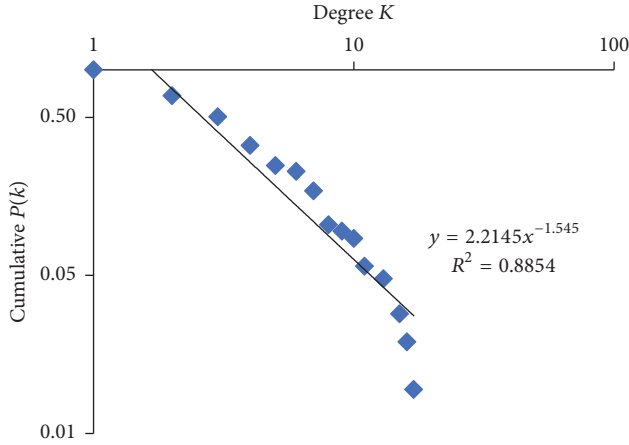


FIGURE 7: Cumulative degree distribution of CMRN.

which network global efficiency of a network G could be calculated by (1) [47, 48], where n refers to the number of vertices and d_{ij} refers to the distance between two vertices.

$$E(G) = \frac{1}{n(n-1)} \sum_{\forall i,j,i \neq j} \frac{1}{d_{ij}}. \quad (1)$$

The effect of risk control of every risk in CMRN can be measured according to the degree of network global efficiency declined. For example, if the network global efficiency decreases by 0.1 after deletion of vertex 8, it means that the effect of risk control of vertex 8 is 0.1. The better the effect is, the greater the risk will be. The 30 most serious risks in CMRN are identified through calculation, and the risk control effects can be shown in Figure 8. It is observed that roof collapse is the most serious risk and controlling roof collapse could help decrease 32.63% of network global efficiency of CMRN, followed by fire (25.96%) and gas concentration exceeding limit (11.22%). However, due to the interaction between roof collapse and gas concentration exceeding limit, the effect of controlling “roof collapse” and “gas concentration exceeding limit” is not equal to 32.63% plus 25.96%, but 44.03% by calculation. Obviously, measuring the effect of risk control can suggest and designate the directions and key points to further safety management. Anyway, controlling several most serious risks is the most appropriate and most effective approach for preventing accident and further promoting the safety management level in the coal mine production.

5. Discussion

Based on the network theory, an analytical framework has been put forward to promote coal mine production safety, which turns out to be feasible and effective. The proposed network modeling method is a powerful and promising tool to analyze risk in various disciplines. It is envisaged that this study can help managerial personnel deeply understand coal mine risk for the sake of developing necessary strategies that can improve safety management in a dynamic operating environment, especially in emergency.

The potential contributions of this study include four aspects. First, it is beneficial to understand the complexity and transitivity of risks in coal mine. The main topology properties and network properties of CMRN are captured and analyzed. Second, it is conducive to enhance the safety performance by controlling original risks and avoiding derivative accidents. Third, this study has the potential benefits in coal mine emergency and relief, which can help managers make decisions in emergency rescue for lightening the casualties and losses. Additionally, network modeling technique is employed in this study, which may offer a promising approach for the analysis of the accident. Also, the application range of network theory will be enlarged.

The main limitation of this research is that the established network model fails to take the vertex weight into consideration. Moreover, the frequency of risks in Table 3 cannot reflect the vertex weight in current study, and it is very difficult to discern and distinguish the different importance for different risks. Therefore, assigning the weight is quite difficult. That may explain why the network model in this study is unweighted. In the future study, more attention should be paid to improve the network model based on more precise understanding of risks in coal mine. Also, how to reduce the risks in coal mine is a significant direction that deserves further research. Meanwhile, a particular failure knowledge database (FKD) would be significant in studying the coal mine accidents, which is the foundation of case based reasoning (CBR) for analyzing hazard and risk.

What is more, several identified risks seem to be general, rather than specific. There are two reasons for this result. First, the risk identification is carried out on the basis of accident data collected from literature and media. If some detailed information is ignored and not recorded during the investigation of these accidents, the unavailable information may affect the accuracy of risk identification. Second, this research is implemented from a holistic perspective. If the identified risks are too specific, finding the common vertex and constructing the coal mine risk network will be difficult.

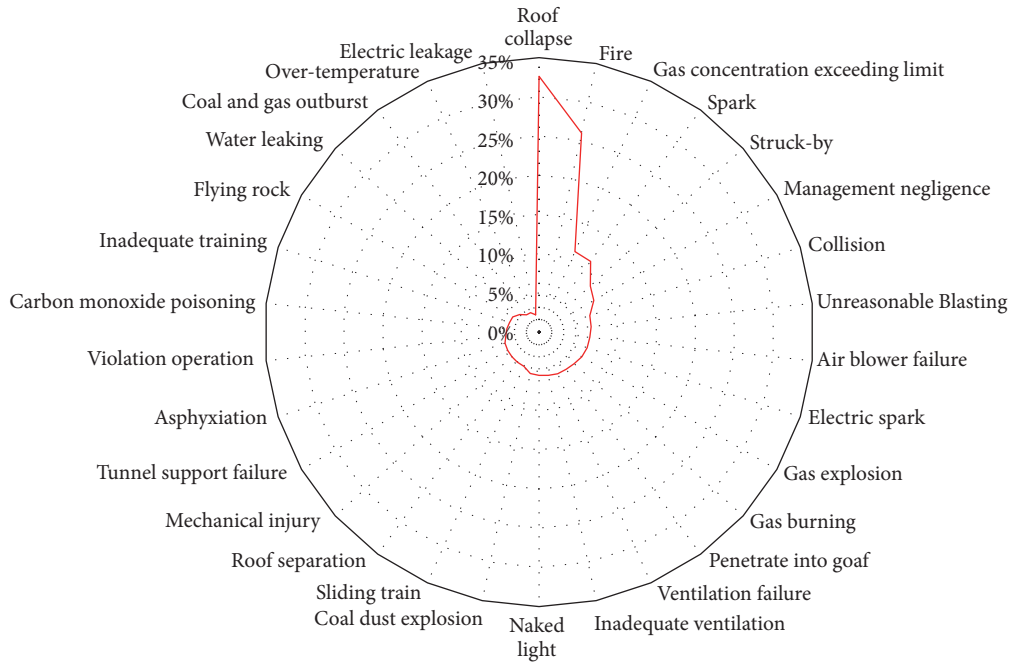


FIGURE 8: The effect of risk control of 30 most serious risks.

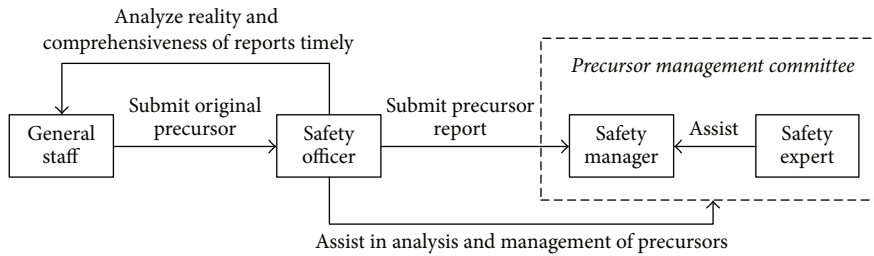


FIGURE 9: The proposed organization structure.

Hence, the similar risks are divided into the same category for the sake of convenience.

Safety researchers assiduously aim to lower the prevalence of accident and raise the safety level. Accident precursor is studied in various industries, and many studies indicate that a series of precursors always occur before the accident [8, 11, 49]. Therefore, lowering precursor frequency is an effective approach to reduce the accident probability [50]. An accident precursor is broadly defined as the “conditions, events, and sequences that precede and lead up to an accident” by the National Academy of Engineering [51]. Even though it is impossible to completely prevent coal mine accident, monitoring and controlling precursory information is a useful and effective approach for safety managers to identify hazards or risks in advance. Also, this can reduce the possibility of accident or alleviate their consequence. Hence, precursor analysis seemingly has huge potential to promote safety management of coal mine production.

According to the results of network analysis, it can be known that some key vertexes play an important role in accident prevention. In practice, precursor can be used to reduce the frequency and probability of these key risks. For example,

the precursor of water leaking mainly includes the following: air turns cold; mist appears on the roadway; and coal wall has water seepage. If the coal miner can pay more attention to these precursors, the water leaking will be reduced to a large extent. Therefore, an organization structure should be proposed to manage and control precursors, as depicted in Figure 9. The real executors of precursor management include general staff, safety manager, and safety expert. General staff should report precursor to the safety officer, and then safety officer submits precursor reports to the safety manager. Furthermore, safety expert assists safety manager to analyze risk as well as factor and propose processing measures. The proposed measures or solutions are executed by general staff and safety officer, and meanwhile, the evaluation of them is implemented by safety manager and safety expert. The coal mine enterprise can set up a committee, mainly including safety manager and safety expert, to deal with precursor management.

6. Conclusion

The accidents in the coal industry have been widely analyzed to promote safety production. By changing the original

method of analyzing a single accident, this research aims to develop an innovative approach of fusing various risks that can explore the full complexity of CMRN based on network theory.

The CMRN is constructed by software *Pajek* based on 135 typical accident chains, which are obtained from 126 typical accidents in coal mine accident database (CMAD). As an unweighted directed network model, CMRN includes 105 vertexes and 194 edges. The network diameter in the CMRN is 7 and the network density of CMRN is 0.178, which indicates that CMRN refers to a relatively sparse network. The value of the average path length in CMRN is 3.0841, suggesting that a risk can transmit to another only in three steps on average. Roof collapse (vertex 68) has the highest degree of 17, which indicates that roof collapse plays a critical role in the accident chain. In general, this type of vertex is regarded as a key point. The vertex betweenness in the CMRN ranges from 0 to 0.059852. Additionally, the roof collapse (vertex 68) has the highest value of vertex betweenness, which means that the maximum number of shortest paths passes through roof collapse (vertex 68). It is a key link in the process of risk spread. Next, fire (vertex 34) and spark (vertex 75) are 0.048486 and 0.020668, respectively. About 55% shortest paths pass through these five highest betweenness vertexes. Effectively controlling roof collapse, fire, spark, gas concentration exceeding limit, and collision could not only increase the network diameter and average path length but also slow down the efficiency of accident propagation and weaken the chain reaction. The vertex clustering coefficient in CMRN ranges from 0 to 0.5. Moreover, the clustering coefficient of CMRN is 0.0623 in CMRN, which denotes that CMRN has a high degree of cliquishness. Besides, CMRN is a relatively small-world network according to its clustering coefficient and average path length, demonstrating that the risk propagation in CMRN is much faster than a random network. CMRN also has the scale-free property because cumulative $P(k)$ follows the power-law. The property indicates that CMRN is robust to random risks to some extent. Furthermore, the effect of risk control is calculated precisely. Overall, roof collapse, fire, and gas concentration exceeding limit are not only three most valuable targets in safety management but also the three most dangerous risks in coal mine production.

Precise calculation of these six parameters and effective risk control are beneficial to capture the complexity and nature of coal mine accident and designate the targets for risk control. Also, the results can help promote coal mine safety management in controlling original risks and preventing derivative accidents. In view of the sequential interrelations among accidents in CMRN, this research may also positive influence the early warning of accidents. In practice, the safety managers should focus more on the identified and valuable targets of risk control and put more resources to help promote safety performance in coal mine production.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors also gratefully acknowledge those who provided data and suggestions. The research described in this paper is supported by National Natural Science Foundation of China (51323004), the Humanities and Social Sciences Youth Foundation of China's Education Ministry (17YJCZH035), the Fundamental Research Funds for the Central Universities (2017QNB13), and Jiangsu Planned Projects for Postdoctoral Research Funds (1701143C).

References

- [1] S. Niu, "Coal mine safety production situation and management strategy," *Management and Engineering*, vol. 14, 78 pages, 2014.
- [2] P. S. Paul, "Predictors of work injury in underground mines - an application of a logistic regression model," *Mining Science and Technology*, vol. 19, no. 3, pp. 282–289, 2009.
- [3] V. V. Khazode, J. Maiti, and P. K. Ray, "A methodology for evaluation and monitoring of recurring hazards in underground coal mining," *Safety Science*, vol. 49, no. 8, pp. 1172–1179, 2011.
- [4] A. Nieto, Y. Gao, L. Grayson, and G. Fu, "A comparative study of coal mine safety performance indicators in China and the USA," *International Journal of Mining and Mineral Engineering*, vol. 5, no. 4, pp. 299–314, 2014.
- [5] Q. Liu, X. Meng, M. Hassall, and X. Li, "Accident-causing mechanism in coal mines based on hazards and polarized management," *Safety Science*, vol. 85, pp. 276–281, 2016.
- [6] S. Mahdevari, K. Shahriar, and A. Esfahanipour, "Human health and safety risks management in underground coal mines using fuzzy TOPSIS," *Science of the Total Environment*, vol. 488–489, no. 1, pp. 85–99, 2014.
- [7] H. Chen, Q. Feng, R. Long, and H. Qi, "Focusing on coal miners' occupational disease issues: A comparative analysis between China and the United States," *Safety Science*, vol. 51, no. 1, pp. 217–222, 2013.
- [8] Z. Zhou, Q. Li, and W. Wu, "Developing a versatile subway construction incident database for safety management," *Journal of Construction Engineering and Management*, vol. 138, no. 10, pp. 1169–1180, 2011.
- [9] J. Santos-Reyes and A. N. Beard, "A systemic analysis of the Edge Hill railway accident," *Accident Analysis & Prevention*, vol. 41, no. 6, pp. 1133–1144, 2009.
- [10] J. K. Wachter and P. L. Yorio, "A system of safety management practices and worker engagement for reducing and preventing accidents: an empirical and theoretical investigation," *Accident Analysis & Prevention*, vol. 68, pp. 117–130, 2014.
- [11] A. Al-shanini, A. Ahmad, and F. Khan, "Accident modelling and analysis in process industries," *Journal of Loss Prevention in the Process Industries*, vol. 32, pp. 319–334, 2014.
- [12] Z. Zhou, J. Irizarry, and Q. Li, "Using network theory to explore the complexity of subway construction accident network (SCAN) for promoting safety management," *Safety Science*, vol. 64, pp. 127–136, 2014.
- [13] P. Andrews-Speed, M. Yang, L. Shen, and S. Cao, "The regulation of China's township and village coal mines: A study of complexity and ineffectiveness," *Journal of Cleaner Production*, vol. 11, no. 2, pp. 185–196, 2003.
- [14] H. Chen, H. Qi, R. Long, and M. Zhang, "Research on 10-year tendency of China coal mine accidents and the characteristics of human factors," *Safety Science*, vol. 50, no. 4, pp. 745–750, 2012.

- [15] H. Chen, Q. Feng, and J. Cao, "Rent-seeking mechanism for safety supervision in the Chinese coal industry based on a tripartite game model," *Energy Policy*, vol. 72, pp. 140–145, 2014.
- [16] N. A. Grazhevska, A. Virchenko, and A. Grazhevska, "The effects of rent-seeking behavior on the efficiency of fiscal policy in Ukraine," *Procedia Economics and Finance*, vol. 27, pp. 274–287, 2015.
- [17] T.-H. Kwon, "Rent and rent-seeking in renewable energy support policies: feed-in tariff vs. renewable portfolio standard," *Renewable & Sustainable Energy Reviews*, vol. 44, pp. 676–681, 2015.
- [18] H. Chen, Q. Feng, D. Zhu, S. Han, and R. Long, "Impact of rent-seeking on productivity in Chinese coal mine safety supervision: a simulation study," *Energy Policy*, vol. 93, pp. 315–329, 2016.
- [19] M. Sari, A. S. Selcuk, C. Karpuz, and H. S. B. Duzgun, "Stochastic modeling of accident risks associated with an underground coal mine in Turkey," *Safety Science*, vol. 47, no. 1, pp. 78–87, 2009.
- [20] C. Qing-gui, L. Kai, L. Ye-jiao, S. Qi-hua, and Z. Jian, "Risk management and workers' safety behavior control in coal mine," *Safety Science*, vol. 50, no. 4, pp. 909–913, 2012.
- [21] I. J. Kowalska, "Risk management in the hard coal mining industry: social and environmental aspects of collieries' liquidation," *Resources Policy*, vol. 41, no. 1, pp. 124–134, 2014.
- [22] A. Badri, S. Nadeau, and A. Gbodossou, "A new practical approach to risk management for underground mining project in Quebec," *Journal of Loss Prevention in the Process Industries*, vol. 26, no. 6, pp. 1145–1158, 2013.
- [23] L. Wang, Y. Wang, Q. Cao, X. Li, J. Li, and X. Wu, "A framework for human error risk analysis of coal mine emergency evacuation in China," *Journal of Loss Prevention in the Process Industries*, vol. 30, no. 1, pp. 113–123, 2014.
- [24] Q.-L. Liu and X.-C. Li, "Modeling and evaluation of the safety control capability of coal mine based on system safety," *Journal of Cleaner Production*, vol. 84, no. 1, pp. 797–802, 2014.
- [25] E. Ghasemi, M. Ataei, K. Shahriar, F. Sereshki, S. E. Jalali, and A. Ramazanzadeh, "Assessment of roof fall risk during retreat mining in room and pillar coal mines," *International Journal of Rock Mechanics and Mining Sciences*, vol. 54, pp. 80–89, 2012.
- [26] L. M. Pejic, J. G. Torrent, E. Querol, and K. Lebecki, "A new simple methodology for evaluation of explosion risk in underground coal mines," *Journal of Loss Prevention in the Process Industries*, vol. 26, no. 6, pp. 1524–1529, 2013.
- [27] A. Bahri Najafi, G. R. Saeedi, and M. A. Ebrahimi Farsangi, "Risk analysis and prediction of out-of-seam dilution in long-wall mining," *International Journal of Rock Mechanics and Mining Sciences*, vol. 70, pp. 115–122, 2014.
- [28] J. Chen, L. Ma, C. Wang, H. Zhang, and M. Ha, "Comprehensive evaluation model for coal mine safety based on uncertain random variables," *Safety Science*, vol. 68, pp. 146–152, 2014.
- [29] D. V. Petrović, M. Tanasijević, V. Milić, N. Lilić, S. Stojadinović, and I. Svrkota, "Risk assessment model of mining equipment failure based on fuzzy logic," *Expert Systems with Applications*, vol. 41, no. 18, pp. 8157–8164, 2014.
- [30] D. R. Lokhande, V. M. S. R. Murthy, V. Vellanky, and B. K. Singh, "Assessment of pot-hole subsidence risk for Indian coal mines," *International Journal of Mining Science and Technology*, vol. 25, no. 2, article no. 467, pp. 185–192, 2015.
- [31] M. Spada and P. Burgherr, "An aftermath analysis of the 2014 coal mine accident in Soma, Turkey: Use of risk performance indicators based on historical experience," *Accident Analysis & Prevention*, vol. 87, pp. 134–140, 2016.
- [32] E. Sun, X. Zhang, and Z. Li, "The internet of things (IOT) and cloud computing (CC) based tailings dam monitoring and pre-alarm system in mines," *Safety Science*, vol. 50, no. 4, pp. 811–815, 2012.
- [33] K. M. Dange and R. T. Patil, "Design of monitoring system for coal mine safety based on MSP430," *International Journal of Engineering Science Invention*, vol. 2, no. 7, pp. 14–19, 2013.
- [34] C. Bo, C. Xin, Z. Zhongyi, Z. Chengwen, and C. Junliang, "Web of things-based remote monitoring system for coal mine safety using wireless sensor network," *International Journal of Distributed Sensor Networks*, vol. 2014, Article ID 323127, 14 pages, 2014.
- [35] Y. Zhang, W. Yang, D. Han, and Y.-I. Kim, "An integrated environment monitoring system for underground coal mines-Wireless Sensor Network subsystem with multi-parameter monitoring," *Sensors*, vol. 14, no. 7, pp. 13149–13170, 2014.
- [36] J. Xu, H. Gao, J. Wu, and Y. Zhang, "Improved safety management system of coal mine based on iris identification and RFID technique," in *Proceedings of the IEEE International Conference on Computer and Communications, ICCCC 2015*, pp. 260–264, China, October 2015.
- [37] W. De Nooy, A. Mrvar, and V. Batagelj, *Exploratory Social Network Analysis with Pajek*, vol. 27, Cambridge University Press, 2011.
- [38] B. W. Wambeke, M. Liu, and S. M. Hsiang, "Using Pajek and centrality analysis to identify a social network of construction trades," *Journal of Construction Engineering and Management*, vol. 138, no. 10, pp. 1192–1201, 2011.
- [39] A. Abbasi, L. Hossain, and L. Leydesdorff, "Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks," *Journal of Informetrics*, vol. 6, no. 3, pp. 403–412, 2012.
- [40] A. M. M. González, B. Dalsgaard, and J. M. Olesen, "Centrality measures and the importance of generalist species in pollination networks," *Ecological Complexity*, vol. 7, no. 1, pp. 36–43, 2010.
- [41] Almaas, E., Kulkarni, R. V., & Stroud, D. (2002). Characterizing the structure of small-world networks. *Physical review letters*, 88(9), 098101.
- [42] S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin, "Catastrophic cascade of failures in interdependent networks," *Nature*, vol. 464, no. 7291, pp. 1025–1028, 2010.
- [43] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.
- [44] A. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.
- [45] S. Ghosh, A. Banerjee, N. Sharma et al., "Statistical analysis of the Indian railway network: a complex network approach," *Acta Physica Polonica B*, vol. 4, no. 2, pp. 123–137, 2011.
- [46] M. E. Newman, "Assortative mixing in networks," *Physical Review Letters*, vol. 89, no. 20, Article ID 208701, 2002.
- [47] P. Crucittia, V. Latorab, M. Marchioric, and A. Rapisarda, "Efficiency of scale-free networks: error and attack tolerance," *Physica A: Statistical Mechanics and its Applications*, vol. 320, pp. 622–642, 2003.
- [48] R. Criado, A. García del Amo, B. Hernández-Bermejo, and M. Romance, "New results on computable efficiency and its stability for complex networks," *Journal of Computational and Applied Mathematics*, vol. 192, no. 1, pp. 59–74, 2006.

- [49] N. Khakzad, F. Khan, and P. Amyotte, "Major accidents (gray swans) likelihood modeling using accident precursors and approximate reasoning," *Risk Analysis*, vol. 35, no. 7, pp. 1336–1347, 2015.
- [50] M. Kyriakidis, R. Hirsch, and A. Majumdar, "Metro railway safety: an analysis of accident precursors," *Safety Science*, vol. 50, no. 7, pp. 1535–1548, 2012.
- [51] J. H. Saleh, E. A. Saltmarsh, F. M. Favarò, and L. Brevault, "Accident precursors, near misses, and warning signs: critical review and formal definitions within the framework of Discrete Event Systems," *Reliability Engineering & System Safety*, vol. 114, no. 1, pp. 148–154, 2013.