Applied Network Science

RESEARCH Open Access

Social networks from dailies: the observer's point of view

José Antonio Motilla-Chávez[1], Edgardo Ugalde[2] and Edgardo Galán-Vásquez[3*]

*Correspondence:
edgardo.galan@iimas.unam.mx

[1] Facultad del Hábitat, Universidad Autónoma de San Luis Potosí, Zona Universitaria Poniente, 79290 San Luis Potosí, Mexico
[2] Instituto de Física, Universidad Autónoma de San Luis Potosí, Zona Universitaria Pedregal, 79295 San Luis Potosí, Mexico
[3] Departamento de Ingeniería de Sistemas Computacionales y Automatización, Instituto de Investigación en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Ciudad Universitaria, 01000 México City, Mexico

Abstract

In this paper, we explore the hypothesis that the social network described by an observer in a written document provides a discernible cross-section of the underlying social network. Specifically, we propose that the subnetwork described by the observer is extracted from large social network with a Barabási-Albert structure, following two key premises: (1) the observer identifies and reports nodes and interactions in his local environment, and (2) the observer records the most relevant nodes—those with a high number of connections—as well as the most significant portions of their local environment. We test this hypothesis using personal dailies written in New Spain and Mexico between 1776 and 1873. We compare the structure of these dailies networks to one depicted by a typical subnetwork of a Barabási-Albert network, obtained following the two premises mentioned above. To support our findings, we compare them to the social networks described in two well-known novels. In this work, we implicitly assume the hypothesis that there is a complete, non-subjective social network, including our observer, with a Barabási-Albert structure. The evolution of this underlying social network can shed light on the dynamics of a society over time. Our approach could give place to indicators of the changes in the structure of this enveloping network. This is especially important since, although our historical sources do not offer sufficient data to reconstruct the processes in their entirety, the approach could supply clues or fragments from which we could understand the structure and dynamics of societies over time.

Keywords: Social network, Barabási's model, NLP, Dailies, Underlying social network

Introduction

Complex networks may represent various natural phenomena, including the World Wide Web, metabolic networks, food web networks, neural networks, communication networks, citation networks, social networks, etc. (Adamic and Huberman 2000). These networks are composed of a collection of nodes that represent system elements and a set of edges that represent interactions between each pair of elements (Barabási and Albert 1999; Newman 2003; Barabasi and Oltvai 2004; Boccaletti et al. 2006).

The research of complex networks has advanced significantly in the previous decade, creating new concepts and metrics to characterize the topology of real networks. As a result, a set of generic characteristics that describe the structure of most natural

networks has been identified (Barabási and Albert 1999; Watts and Strogatz 1998; Shen-Orr et al. 2002; Palla et al. 2005).

Among the most common topological measures devised to characterize a network, we have the degree of a node k, which is the number of interactions it has with other nodes. This quantity has great relevance because it describes the network's connectivity (Newman 2003). The degree distribution $P(k)$ gives the probability that a node chosen uniformly at random has a degree k. It has been widely verified that a huge variety of real-world networks exhibit a power law distribution $P(k) \sim Ak^{-\gamma}$ with an exponent taking a value between 2 and 3 (Barabási and Albert 1999).

A power law distribution reflects that many nodes have low connectivity while a few are highly connected. Barabási and Albert have proposed a mechanism to explain the scale-free behavior observed in these networks (Barabási and Albert 1999). They assume that the network is the result of a process where the number of nodes increases over time in such a way that the new nodes establish links with the old ones following a preferential attachment rule: a connection is randomly formed between a newly created node and a previously present one, with a probability proportional to the degree of the latter (Barabási and Albert 1999). It has to be noted that the importance of this kind of distribution, and its emergence as a result of a dynamic process with cumulative advantage, was discovered and studied by de Solla Price (de Solla Price 1965, 1976) several decades before Barabási and Albert.

Social networks represent the interactions between individuals within the same social group, whether in real-world or fictional settings. In these networks, nodes correspond to agents within a particular class-most commonly individuals recognized as part of the same group-and edges signify the relationships between them (e.g., family ties, friendships, or professional connections, etc.) (Tabassum et al. 2018). The study of social networks emerges as a tool for analyzing the underlying structures that reveal themselves to share some universal characteristics (Fronczak 2014).

A recent approach was developed to study networks extracted from literary texts. To build them, different strategies were considered, such as the reconstruction of networks from British novels and series (Elson et al. 2010), using tools such as ALCIDE for automatic reconstruction (Diesner 2013) and approximations from natural language text data (Moretti et al. 2014).

As Labatut and Bost refer (Labatut and Bost 2019), recent research dedicated to the analysis of literary fiction texts has focused on studying the structure and interaction of the characters throughout a story, which they consider to be the central and articulating axis. They group these works into three broad fields: narrative analysis, in which, based on the analysis of small corpora, researchers manually extract character networks (Moretti 2011; Xanthos et al. 2016; Rochat and Triclot 2017; Bounegru et al. 2017); secondly, approaches that rely on the complexity paradigm, through the use and development of tools for corpus analysis (Sudhahar and Cristianini 2013; Tan et al. 2014); thirdly, analysis that uses Artificial Intelligence tools to extract various elements from the text automatically (Jung et al. 2013; Ardanuy and Sporleder 2014).

In this work, the focus is put on documents written from the point of view of an observer. An interesting precedent is the work of Espinal et al. (Espinal-Enríquez et al. 2015), where the authors reconstruct and analyze the topological structure implicit in

the book *Los señores del narco* by Anabel Hernández. There, the interactions between different criminal organizations are reported from the point of view of the person who compiles the information. Espinal et al. construct a network by considering all the characters mentioned in Hernández's book and establishing a link between two of them if they co-occur at a certain distance in the text.

Additionally, in a previous study Motilla-Chavez et al. (2021), we analyzed the social network contained in a diary, which allowed us to observe the transitions of government during the War of Reform in México. Thought this, we gained insights into the dynamics of a particular sector of society during wartime. From this, we hypothesize that texts written from the point of view of an observer describe an underlying social network, which is extracted from a social network with a Barabási-Albert structure. The study of a network described from the point of view of an observer may supply a way to deduce the salient topological characteristics of the enveloping social network. This will make it possible to understand certain aspects of a society's dynamics at a given time. This work constitutes a first step in this direction.

Methodology

Datasets

The datasets were obtained from three main sources. The first source comprises three diary texts written in New Spain and Mexico between 1776 and 1873. These include: *Juan Vildósola's diary*, a diary written by a Catholic seminarian in San Luis Potosí that details his day-to-day life during a war conflict in Mexico. This diary consists of 238 pages and reports events involving 914 people. The second diary, *Diario de sucesos notables de México del alabardero José Gómez (1776–1798)*, chronicles the notable events of José Gómez's life during the period of New Spain (now Mexico). It contains 297 pages and reports events involving 978 people. The third diary, *Diary of Don Agustín Soberón Sagredo (1819–1873)*, documents events in the state of San Luis Potosí, México, spanning 203 pages and involving 2246 people.

The second dataset corresponds to two novels: *Ulysses*, written by James Joyce and published in 1922, consists of 732 pages and includes 273 characters. The second novel, *Chronicle of a Death Foretold*, written by Gabriel García Márquez in 1981, consists of 122 pages and features 55 characters. These two datasets represent sources written from the perspective of a narrator. In *Ulysses*, the narrator describes events and experiences drawn from the author's life, whereas *Chronicle of a Death Foretold* is a work of fiction narrated from a third-person perspective.

A third set of networks was selected from previously published social networks, which include a sample of Facebook's social network, a network of scientific collaboration, and *Los señores del narco's network*. The Facebook network is a compilation of interactions among users in the New Orleans Facebook network. This is an undirected network where the nodes represent users and the edges represent friendships between them. It contains 63,731 nodes and 817,035 edges (Viswanath et al. 2009). The scientific collaboration network is a collection of collaborations among authors of papers in the general relativity and quantum cosmology categories. This is also an undirected network, where an interaction is defined if author i co-authored a paper with author j, resulting in an edge between i and j. It consists of 5,242 nodes and 14,496 edges (Leskovec et al.

2007). The Los señores del narco's networks is a social network constructed through text mining from the book *Los señores del narco*. This network was built by identifying the position of each name in the text, and an interaction was annotated if two names co-occurred within 200 bytes. It comprises 1,037 nodes and 6,405 edges (Espinal-Enríquez et al. 2015) (Supplementary Table S1).

Network reconstruction

Each text was digitized using Czur ET24 Pro scanner, which includes a powerful OCR (optical character recognition) software developed by the company ABBYY. Subsequently, we manually curated the texts to correct characters, punctuation, and structure errors. This process also involved the normalization of the text which consists in the transformation of the text in a single form, replacing pronouns with the full names of the individuals mentioned. After this, we manually constructed a name index for each text.

Next, each text and its corresponding name index were converted to uppercase letters, this was done to prevent case-sensitive errors. The texts were tokenized with Python, version 3.7, using the Natural Language Toolkit (NLTK) library (Bird et al. 2009). Each token is treated as an event, and corresponds to a single phrase within the text.

To reconstruct the underlying social networks, each index in $V = \{1, 2, \ldots, N\}$ was searched inside each token. An interaction was considered to exist if two indexes appear in the same sentence, i.e., if i and j co-occurrence in the same token, the link $i \leftrightarrow j$ is included in the network (Fig. 1).

Topological analysis

A network $G = (V, E)$ is formed by a set V of nodes, which in our case represent people, and a set E of edges (links between nodes) that indicates the co-occurrence of people in the same sentence. To characterize the networks' topology, diverse metrics

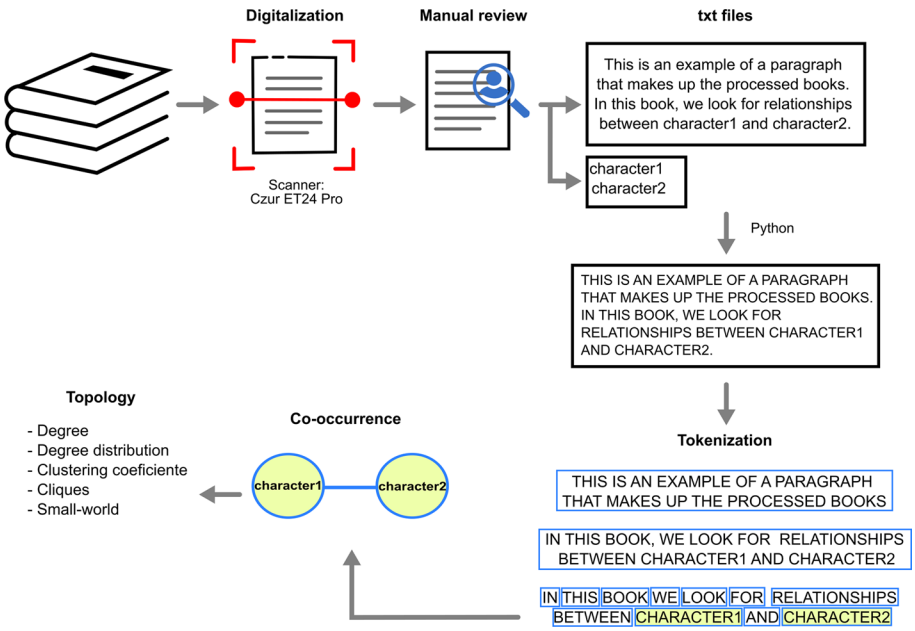

Fig. 1 Workflow for building underlying social networks

such as node degree, degree distributions, clustering coefficient, community structure, connected component, and the small-world index (Newman 2003; Palla et al. 2005) were computed.

We denote by k_i the degree of the node i, calculated as the number of edges incident on this node. In an undirected network, the number of edges E can be expressed as the sum of the node degrees:

$$E = \frac{1}{2}\sum_{i=1}^{N} k_i$$

The factor of 1/2 corrects for the fact that in undirected networks, each link is counted twice.

The degree distribution $P(k)$ provides the probability that a randomly selected node in the network has a degree k. The degree distribution was calculated considering the relative of occurrence for every degree k, such that:

$$\sum_{K=1}^{\infty} P(k) = 1$$

The clustering coefficient measures the degree to which the neighbors of a given node are interconnected. For a node i of degree k_i, the local clustering coefficient is defined as:

$$C_i = \frac{2E_i}{k_i(k_i - 1)}$$

where E_i represents the number of interactions between the k_i neighbors of node i. The coefficient C_i can be interpreted as the probability that two neighbors of a node are also neighbors with each other.

In a given society, there are diverse groupings, such as family, labor organizations, friend circles, collectives, social clubs, etc. To detect these communities inside the network, we calculated cliques. A clique O is a subset of nodes connected to each other. Hence, the subgraph $G(O)$, induced by those nodes, is completely connected.

The small-world property describes the phenomenon where, for each pair of nodes, the shortest path connecting them is relatively short compared to the total number of nodes in the network. This property facilitates the rapid diffusion of information through the network, much faster than in networks lacking this characteristic (Newman 2003; Wang and Chen 2003). The small-world property (σ) is determined by:

$$\sigma = \frac{C/C_r}{L/L_r} > 1,$$

where C is the clustering coefficient of the analyzed network, while C_r is the clustering coefficient expected from a random network with the same connectivity; L represents the length of the largest path among all the shortest paths connecting two nodes and is known as the diameter of the network, while L_r is the diameter of a random network with the same connectivity. This coefficient also indicates that the network tends to contain cliques, fully connected subnetworks grouping nodes with similar characteristics.

Model

To understand the generation of such underlying networks from an observer's point of view, we developed a growth model implemented in Python 3.7, following the guidelines provided below:

1. We begin with the idea that the observer samples a portion of a social network that exhibits a scale-free structure. To simulate this, we recreated a social network using the Barabási-Albert preferential attachment model (Barabási and Albert 1999), with the Package NetworkX and function *barabasi_albert_graph*. It started with a seed of three nodes and adding three edges from each new node to preexisting nodes until a total of 10,000 nodes was achieved. The degree distribution $P(k)$ was computed once the network was built, verifying a power-law behavior $P(k) \sim A\, k^{-\gamma}$.
2. Next, we determine the sample size by identifying the number of characters that co-occur within the tokens of the previously analyzed dailies. We found that the distribution of characters with the tokens follows an exponential distribution, as described by the following equation:

$$f(x;\beta) = \frac{1}{\beta} e^{\frac{-x}{\beta}}$$

where x represents the number of characters in each token (sentence), and $\beta = 3$ is the parameter of the distribution.
3. Then, we calculate the degree for each node in the Barabási-Albert network and select a particular node with low connectivity, which we designate as the observer. From this observer node, we identify its nearest neighbor based on the shortest path between pairs of nodes and retain those within a maximum distance of five steps, defining this as the local neighborhood.
4. Additionally, we assume that the observer reports two types of events in the dailies. The first involves describing characters from the local environment, while the second involves reporting interactions between main characters in the original network, i.e., highly connected nodes in the Barabási-Albert social network. Each event corresponds to a small subnetwork sampled as follows:

   (a) For the first type of event, where the reporter writes about the local environment, we assign a probability of 0.2. In this case, the token of size x is populated with nodes from the observer's local neighborhood, which are randomly selected. The 0.2 proportion is indicative and was chosen to align well with the networks extracted from the dailies. This proportion reflects the difficulty of estimating the exact share of events concerning the observer's local network, especially in historical texts considered in this study.
   (b) For the second type of event, where the reporter writes about prominent individuals, we assign a probability of 0.8. In this case, a node is selected based on its degree, so that highly connected nodes have a greater probability of being chosen. Once selected, its neighbors, at varying levels, are included in the event until the required number of characters is met. The neighbors are chosen

through random selection with a preference based on their degree (See Pseudocode in Fig. S1).

Results

To analyze the existence and topology of underlying social networks in texts, we selected a set of three digitized texts written from the perspective of a narrator who reported his daily life in a diary. The events reported in these texts are between 1776 and 1873 in New Spain and Mexico.

Each text was processed through text mining to reconstruct its underlying social network. For this, the text was tokenized according to its sentences, and subsequently, each pair of characters from the name index was searched in each of the tokens. An interaction between two characters is defined if these two characters co-occur within the same sentence. Finally, these relationships were integrated into an undirected network of characters.

Juan Vildosola's network was built from a daily that chronicles events in San Luis Potosí during Mexico's reform war, in which the author, a catholic seminarian, describes the day-to-day life of a conflicted community. There are 584 nodes and 2228 edges between the characters in this network (Motilla-Chavez et al. 2021). Jose Gomez's network was built from the *Diario de sucesos notables de México del alabardero José Gómez*. It chronicles life in the new Spain (now México) from 1776 to 1798. This network has 584 nodes and 2228 edges between the characters. Finally, the *Diario de don Agustín Soberón Sagredo* chronicles key events in San Luis Potosí, México, from 1858 to 1873. There are 627 nodes in this network, with 2501 connections between the characters (Agustin Soberon's network). While our strategy of underlying social network reconstruction identifies characters that co-occur within tokens, it does not account for individually mentioned characters-those who are reported in the text but do not interact with other characters in the same sentence-resulting in the exclusion of some characters from the network.

Topology in dailies

We used the graph theoretical approach to characterize these networks' structure. In this context, the degree k of a node is defined as the number of interactions that it has with other nodes, while the degree distribution gives the probability $P(k)$ of finding a node with degree k (Albert 2005). This measure quantifies the diversity of degrees in a network (Watts 2004). Many natural networks have a degree distribution close to a power law $P(k) \sim A\,k^{-\lambda}$, corresponding to a network with no characteristic degree, where few nodes are highly connected (they are called hubs). In contrast, most have low connectivity (Barabasi and Oltvai 2004). In the case of networks derived from the dailies, we found that the degree distribution fits a power-law more accurately, with $\lambda < 2$ (Fig. 2), significantly reducing the mean square error for each distribution compared to the Poisson distribution (see Fig. S1).

The clustering coefficient, C, indicates the probability that two nodes with a common neighbor in a graph are also interconnected; that is, the clustering coefficient quantifies in what proportion the local neighborhood of a node forms a clique. We find

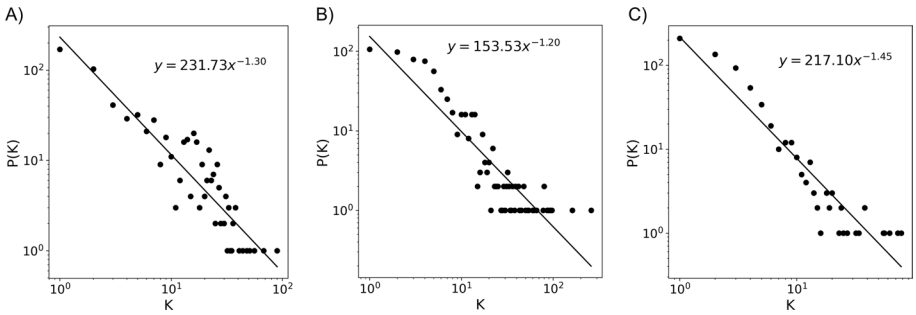

Fig. 2 Degree distribution of the dailies. (**A**) Juan Vildosola's network, (**B**) Agustin Soberon's network, and (**C**) Jose Gomez's network. *P*(*k*) tells us the probability that a randomly chosen node will have degree *K*. The points reflect the frequency of specific degree found in each network

Table 1 Main topological measures

Network	Nodes	Edges	λ	C	σ	Cliques
Facebook	63,731	817,035	2.05	0.221	> 1	1,539,038
Collaboration	5242	14,496	2.04	0.530	> 1	3906
Los señores del narco	1037	6405	1.27	0.623	2.634	1459
Juan Vildosola	634	2555	1.3	0.542	5.4	456
Jose Gomez	631	1318	1.45	0.464	2.6	539
Agustin Soberon	649	2833	1.20	0.624	1.810	1047
Ulysses	110	321	0.89	0.408	2.295	83
Chronicle of a death foretold	47	159	0.758	0.513	1.080	64

$C = 0.533, 0.382$, and 0.382 for Juan Vildosola's, Jose Gomez's, and Agustin Soberon's networks, respectively. This indicates that networks tend to create tightly-knit groups.

To determine if the diaries' networks have a topology similar to other social networks, we compared the structure of our networks to three previously published networks: a subnetwork of Facebook belonging to New Orleans users (Viswanath et al. 2009), and a collaboration network that is a collection of collaborations between authors papers in the category of General relativity and quantum cosmology (Leskovec et al. 2007), and a network built from the book *Los señores del narco* (Espinal-Enríquez et al. 2015).

We identified that our diaries' networks have similar topological characteristics to other previously published social networks (Table 1). However, we identified a difference in the exponent λ of published social networks such as Facebook and science collaboration, where λ has values between 2 and 3 (Fig. 3A,B), while for the networks extracted from the book *Los señores del narco*, λ has a value between 1.2 and 1.3 (Fig. 3C).

The exponent λ provides information about the network structure. Networks with $\lambda > 3$ lack many of the properties of a scale-free network, which are present for $2 \leq \lambda \leq 3$, where a hierarchy in the nodes' degree appears. For $\lambda = 2$, the node with the highest degree influences a large fraction of the network (Barabasi and Oltvai 2004). As shown below, an exponent $\lambda < 2$ is consistent with a subnetwork of a Barabási-Albert network extracted from the point of view of an observer node.

Although scale-free networks with $\lambda > 2$ are more widely present in the real world, networks with $\lambda < 2$ have also been identified, such as Wiki vote, Political blogs,

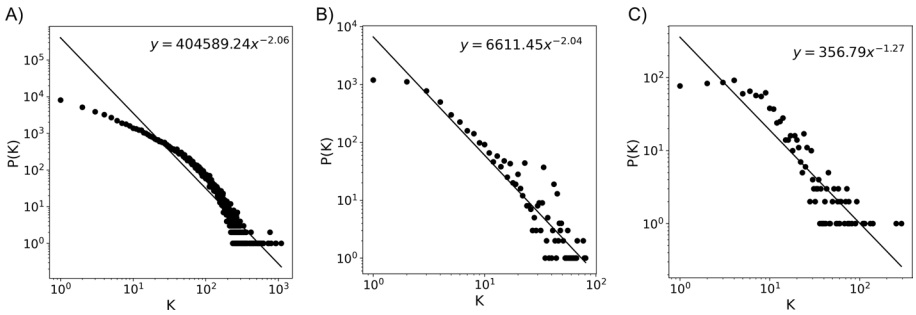

Fig. 3 Degree distribution of previously published networks. (**A**) Facebook network, (**B**) Collaborations network, and (**C**) Los señores del narco network. *P*(*k*) tells us the probability that a randomly chosen node will have degree *K*. The points reflect the frequency of specific degree found in each network

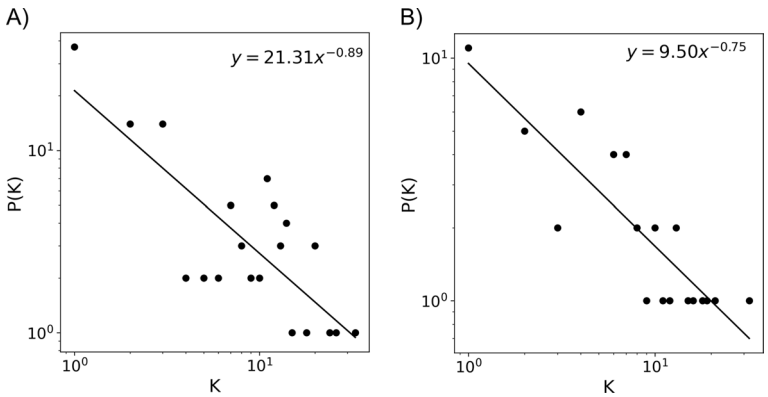

Fig. 4 Degree distribution of the fiction novels. (**A**) *Ulysses* network and (**B**) *Chronicle of a Death Foretold* network. *P*(*k*) tells us the probability that a randomly chosen node will have degree *K*. The points reflect the frequency of specific degree found in each network

Slaahdot, Epinions, Emails, Gnuyella and Dependency. These were first explored by Seyed-Allaei et al. (2006), and the main property is that the number of links grows faster than the number of nodes. Furthermore, the maximum degree and the average degree grow rapidly with the size of the network, and there are many hub nodes (Seyed-Allaei et al. 2006; Li et al. 2018).

Networks from fiction novels

To evidence the realistic nature of the social networks underlying the networks described in dailies, we compare their structure to those depicted in works of fiction. These networks were reconstructed with the same strategy; we use James Joyce's *Ulysses* and Gabriel García Márquez's *Chronicle of a Death Foretold*.

We identify that the networks of fictional texts have a structure completely different from the networks we extracted from diaries (Fig. 4). The latter are networks of a few nodes strongly interacting with one another, resulting in a high proportion of nodes being highly connected. This is reflected by a degree distribution not very close to a power law, and if fitted to one, with an exponent $\lambda < 1$. However, the network's topology does not follow a random distribution of nodes (Fig. S2), as certain characters exhibit significantly higher degrees of connectivity compared to others.

Topology of the model's network

Several models are reported in the literature whose degree's distribution is a power-law with exponent $\lambda < 2$. Li et al. explored an insertion-deletion-compensation model giving place to a scale-free complex network. Using a mean-field approach, they show that the degree distribution is a power law, with exponents ranging from 1 to 3 (Li and Tang 2020). In related work, Li and Tang proposed a model whose construction is governed by a Poisson process and a selection process with probability p, giving place to a power law with an exponent $\lambda < 2$, which depends on the parameters of the Poisson and selection processes (Li et al. 2018). It is worth mentioning the work by Hill and Braha (Hill and Braha 2010), where a persistent random walk, over the underlying Barabási-Albert network, generates scale-free subnetworks with an exponent $\lambda < 2$. Their construction aims to model dynamical centrality observed in some real communication networks, generating connected subgraphs evolving on time.

Here, we adopt a different approach to constructing the social networks of diaries, based on the hypothesis they are subnetworks extracted by an observer from a more extensive social network with a Barabási-Albert structure in such a way that 1) the observer includes nodes and interactions in his local environment, and, 2) he registers the most relevant nodes and a significant part of their local environment. For this, the distribution of the characters in the phrases of each diary was identified, finding an exponential distribution with an exponential rate equal to three. This indicates that most sentences have a co-occurrence of two characters, but there are sentences with more (Fig. 5A).

Several practical considerations were taken into account for the implementation of the models. First, the enveloping network from which the observer extracts a subnetwork is assumed to be a Barabási-Albert network with a seed equal to three, a number of edges to attach from a new node to existing nodes equal to three, and a final number of nodes equal to 10,000 (Fig. 5B). Secondly, the number of characters selected in each reported event follows the previously identified exponential distribution. Furthermore, as mentioned in Subsection 2.4 above, the selection of the characters for each event is done by considering two types of characters: with probability 0.2, an event is completed using the observer's neighbors, and with probability 0.8, a node is chosen in the neighborhood of a highly connected node based on its degree, which indicates that most events are related to the information involving important people (Fig. 5C). We show that the model reproduces the topological characteristics of the networks described in the dailies (Table 2 and Fig. 5D).

Comparison to Erdös-Rényi's enveloping

To discard the fact that the enveloping networks from which the observers extract the dailies networks is a homogeneous random network, we perform the same procedure over Erdös-Rényi networks with 10,000 nodes and 29,991 edges that correspond to a Barabási-Albert network of the same size (Fig. 6A). The exponential distribution of event sizes and the sampling protocol were performed as in the case of the Barábasi-Albert enveloping network (Fig. 6B).

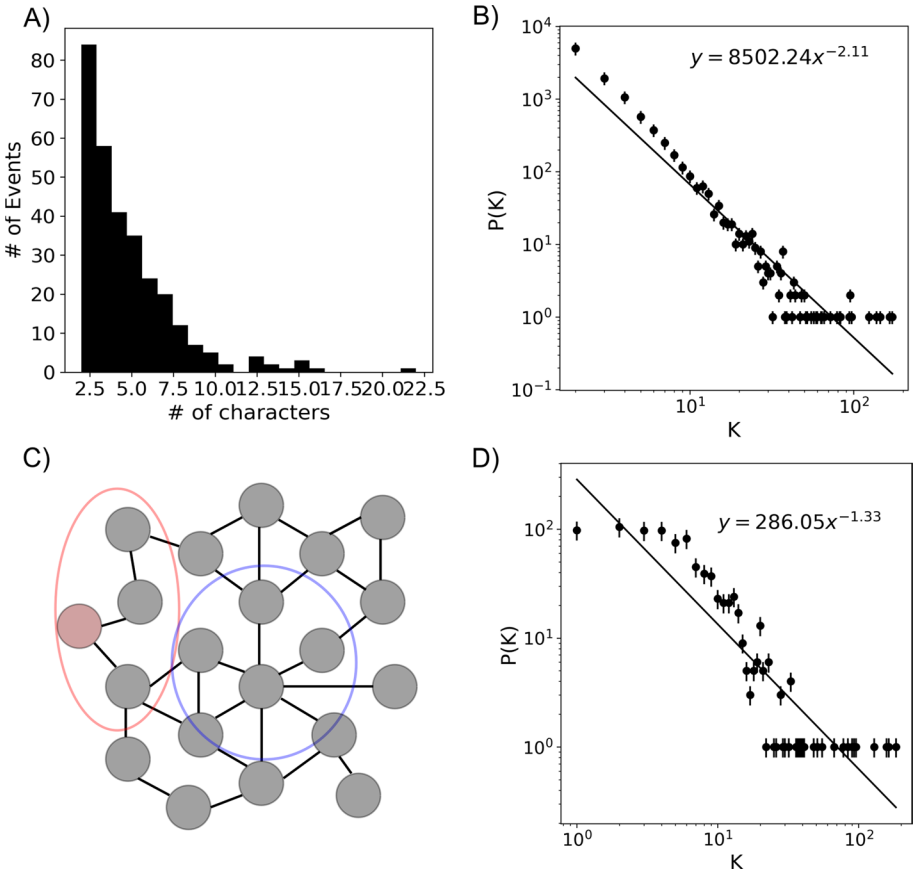

Fig. 5 Scheme of the model. (**A**) Distribution of characters in events. (**B**) Degree distribution of Barabási-Albert network, (**C**) Underlying social networks of dailies. The node-red is labeled as the observer, and the elements in the red circle represent his neighborhood, while the blue circle represents events of highly connected characters, (**D**) Degree distribution of the model

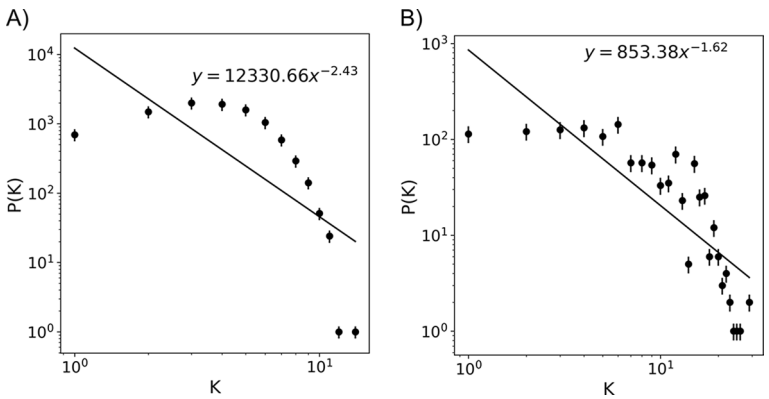

Fig. 6 Random networks. (**A**) Degree distribution of Erdös-Rényi network, and (**B**) Degree distribution of the extracted

We compute, as well, the clustering coefficient C, the small word coefficient σ, the number of cliques, connected components, and the fraction of nodes in the giant component for the subnetwork obtained from the unstructured, homogeneous

Table 2 Topological measurements of the models

Network	λ	C	σ	Cliques	Connected component	Fraction of nodes in giant component
Juan Vildosola	1.34	0.533	> 1	413	46	0.79
Jose Gomez	1.25	0.382	> 1	247	42	0.71
Agustin Soberon	1.20	0.630	> 1	843	14	0.94
Model BA [1]	1.38 ± 0.001 [3]	0.79 ± 0.000 [3]	> 1	412.02 ± 0.92 [3]	72 ± 0.26 [3]	0.72 ± 0.001 [3]
Model ER [2]	1.78 ± 0.010 [3]	0.83 ± 0.001 [3]	> 1	301.37 ± 0.332 [3]	262 ± 1.38 [3]	0.06 ± 0.004 [3]

[1] Network model based on Barabási-Albert networks

[2] Network model based on Erdös-Rényi networks

[3] This is an average value of 1000 runs

random network. This is shown in Table 2, where we compared to Juan Vidosola's, Jose Gomez's, and Agustin Soberon's networks.

Although the clustering coefficient and the number of cliques coincide in the case of both extracted subnetworks, the degree distribution, the number of connected components, and the fraction of nodes in the giant component notably differ, being more accurately described by a power law in the case of the subnetwork extracted from the Barabási-Albert. Nevertheless, it can be observed that dailies' networks exhibit slightly weaker connectivity than the one observed in the model. Despite this connectivity discrepancy, it is clear that Barabasi-Albert's is a more accurate model for the social network underlying the dailies than Erdös-Rényi.

Conclusion

In this study, we proposed that documents written from the perspective of an observer contain an underlying social network extracted from a complex, enveloping social network from which the author selects characters to integrate the reported events. The results show that the global structure of these networks differs from a conventional social network by having a scale-free structure with a $\lambda < 2$, which is consistent with the small-world properties and the presence of a high number of highly-connected nodes. This is consistent with what has been observed in real networks with $\lambda < 2$.

Although different models for scale-free networks with an exponent $\lambda < 2$ have been proposed, our model distinguishes itself from the fact that it implements the idea of a subnetwork extracted by a scarcely connected observer from a Barabási-Albert's social network. This modeling implements the construction of those networks as a compilation of characters' neighborhoods appearing in the accounted events.

From a social point of view, identifying these underlying social networks can indicate a society's dynamics at a given time. Although the available sources do not allow us to fully reconstruct the interactions of a complete social universe, with a structure close to a Barabási-Albert model, they allow us to identify important elements within the social context. This becomes more relevant in historical texts, where it is difficult to reconstruct the processes beyond what archival sources and the narratives prepared by historians report. Using this model allows us to realize a structural analysis of a society that transcends the mere narrative dimension and opens the possibility of a deeper and more reflective understanding of historical processes.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s41109-024-00681-9.

Supplementary file 1. Fig. S1. Pseudocode of the underlying social network generation model.

Supplementary file 2. Fig. S2. Mean square error in fitting the degree distribution using Poisson distribution and power law.

Supplementary file 3. Table S1. Underlying social networks.

Author contributions

Motilla-Chávez JA: conceptualization, funding acquisition, investigation, writing—review and editing. Ugalde E: investigation, writing—review and editing. Galán-Vásquez E: Formal analysis, methodology, visualization, writing—review and editing.

Funding

This work was supported by the grant "Ciencia de Frontera 2023" CF-2023-G-941 of the Consejo Nacional de Humanidades Ciencias y Tecnologías (Conahcyt) de México, and by UNAM-PAPIIT (IA207423).

Availability of data and materials

No datasets were generated or analysed during the current study.

Declarations

Competing interests

The authors declare no competing interests.

Received: 19 July 2024 Accepted: 21 October 2024
Published online: 08 November 2024

References

Adamic LA, Huberman BA (2000) Power-law distribution of the world wide web. Science 287(5461):2115–2115
Albert R (2005) Scale-free networks in cell biology. J Cell Sci 118(21):4947–4957
Ardanuy MC, Sporleder C (2014) Structure-based clustering of novels. In Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL) (pp. 31-39)
Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5(2):101–113
Barabási AL, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512
Bird S, Klein E, Loper E (2009) Natural language processing with Python: analyzing text with the natural language toolkit. "O'Reilly Media, Inc."
Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU (2006) Complex networks: structure and dynamics. Phys Rep 424(4–5):175–308
Bounegru L, Venturini T, Gray J, Jacomy M (2017) Narrating networks: exploring the affordances of networks as storytelling devices in journalism. Digit J 5(6):699–730
Diesner J (2013) From texts to networks: Detecting and managing the impact of methodological choices for extracting network data from text data. KI-Künstliche Intelligenz 27:75–78
Elson DK, McKeown K, Dames NJ (2010) Extracting social networks from literary fiction. Columbia University 1–10
Espinal-Enríquez J, Siqueiros-García JM, García-Herrera R, Alcalá-Corona SA (2015) A literature-based approach to a narco-network. In Social Informatics: SocInfo 2014 International Workshops, Barcelona, Spain, November 11, 2014, Revised Selected Papers 6 (pp. 97-101). Springer International Publishing
Fronczak P (2014) Scale-free nature of social networks. Springer, New York
Hill SA, Braha D (2010) Dynamic model of time-dependent complex networks. Phys Rev E 82:046105
Jung JJ, You E, Park SB (2013) Emotion-based character clustering for managing story-based contents: a cinemetric analysis. Multimedia Tools and Applications 65:29–45
Labatut V, Bost X (2019) Extraction and analysis of fictional character networks: a survey. ACM Comput Surv (CSUR) 52(5):1–40
Leskovec J, Kleinberg J, Faloutsos C (2007) Graph evolution: Densification and shrinking diameters. ACM transactions on Knowledge Discovery from Data (TKDD), 1(1), 2-es
Li Z, Tang X (2020) A study on scale free social network evolution model with degree exponent < 2. J Syst Sci Complexity 33(1):87–99
Li J, Zhou S, Li X, Li X (2018) An insertion-deletion-compensation model with Poisson process for scale-free networks. Futur Gener Comput Syst 83:425–430
Moretti G, Tonelli S, Menini S, Sprugnoli R (2014) ALCIDE: An online platform for the Analysis of Language and Content In a Digital Environment. In Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014 & and of the Fourth International Workshop EVALITA 2014: 9-11 December 2014, Pisa (pp. 270-274). Pisa University Press
Moretti F (2011) Network theory, plot analysis. New Left Rev 68

Motilla-Chavez J, Ugalde E, Galán-Vásquez E (2021) Análisis de un diario de la Guerra de Reforma (1858-1860) en San Luis Potosí, México, por medio de la teoría de redes: una propuesta metodológica para el análisis de textos histórico. In: Violencia, representaciones y estrategias. La guerra y sus efectos en México, Colombia y Guatemala, siglos XVI-XX. COLMEX

Newman ME (2003) The structure and function of complex networks. SIAM Rev 45(2):167–256

Palla G, Derényi I, Farkas I, Vicsek T (2005) Uncovering the overlapping community structure of complex networks in nature and society. Nature 435(7043):814–818

Rochat Y, Triclot M (2017) Les réseaux de personnages de science-fiction: échantillons de lectures intermédiaires. Revue d'études sur la science-fiction, ReS Futurae, p 10

Seyed-Allaei H, Bianconi G, Marsili M (2006) Scale-free networks with an exponent less than two. Phys Rev E-Stat Nonlinear Soft Matter Phys 73(4):046113

Shen-Orr SS, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. Nat Genet 31(1):64–68

de Solla Price DJ (1965) Network of scientific papers. Science 149:510–515

de Solla Price DJ (1976) A general theory of bibliometric and other cumulative advantage processes. J Am Soc Inf Sci 27(5):292–306

Sudhahar S, Cristianini N (2013) Automated analysis of narrative content for digital humanities. Int J Adv Comput Sci 3(9):440–447

Tabassum S, Pereira FS, Fernandes S, Gama J (2018) Social network analysis: an overview. Wiley Interdiscipl Rev: Data Min Knowl Discov 8(5):e1256

Tan MS, Ujum EA, Ratnavelu K (2014) A character network study of two Sci-Fi TV series. In AIP Conference Proceedings (Vol. 1588, No. 1, pp. 246-251). American Institute of Physics

Viswanath B, Mislove A, Cha M, Gummadi KP (2009) On the evolution of user interaction in Facebook. In Proceedings of the 2nd ACM workshop on Online social networks (pp. 37-42)

Wang XF, Chen G (2003) Complex networks: small-world, scale-free and beyond. IEEE Circuits Syst Mag 3(1):6–20

Watts DJ (2004) The "new" science of networks. Annu Rev Sociol 30(1):243–270

Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393(6684):440–442

Xanthos A, Pante I, Rochat Y, Grandjean M (2016) Visualising the dynamics of character networks. In Digital Humanities (pp. 417–419)

Publisher's Note