Physica A 500 (2018) 23–39

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Stylized facts in social networks: Community-based static modeling

Hang-Hyun Jo [a,b,c,*], Yohsuke Murase [d], János Török [e,f], János Kertész [f,e,c], Kimmo Kaski [c]

[a] Asia Pacific Center for Theoretical Physics, Pohang 37673, Republic of Korea
[b] Department of Physics, Pohang University of Science and Technology, Pohang 37673, Republic of Korea
[c] Department of Computer Science, Aalto University School of Science, Espoo FI 00076, Finland
[d] RIKEN Advanced Institute for Computational Science, Kobe, Hyogo 650-0047, Japan
[e] Department of Theoretical Physics, Budapest University of Technology and Economics, Budapest H-1111, Hungary
[f] Center for Network Science, Central European University, Budapest H-1051, Hungary

ARTICLE INFO

Article history:
Received 7 August 2017
Available online 17 February 2018

ABSTRACT

The past analyses of datasets of social networks have enabled us to make empirical findings of a number of aspects of human society, which are commonly featured as stylized facts of social networks, such as broad distributions of network quantities, existence of communities, assortative mixing, and intensity-topology correlations. Since the understanding of the structure of these complex social networks is far from complete, for deeper insight into human society more comprehensive datasets and modeling of the stylized facts are needed. Although the existing dynamical and static models can generate some stylized facts, here we take an alternative approach by devising a community-based static model with heterogeneous community sizes and larger communities having smaller link density and weight. With these few assumptions we are able to generate realistic social networks that show most stylized facts for a wide range of parameters, as demonstrated numerically and analytically. Since our community-based static model is simple to implement and easily scalable, it can be used as a reference system, benchmark, or testbed for further applications.

1. Introduction

Characterizing the social networks is of crucial importance to understand various collective dynamics taking place in them [1–3], as exemplified by disease spreading and diffusion of innovation and opinions. In recent years, the characterization of social networks in the unprecedented detail has become possible because of the availability of a number of large-scale digital datasets, e.g., face-to-face interactions [4–6], emails [7,8], mobile phone communication [9,10], online forums [11,12], Social Networking Services (SNSs) like Facebook [13] and Twitter [14], and even massive multiplayer online games [15,16]. However, these datasets capture only a part of the entire social network, implying that any conclusions derived from such datasets cannot be simply extrapolated to the whole society. Here the entire social network indicates a comprehensive picture of human social relationships with complex community structure due to today's multiple communication channels, and can be called a multi-channel weighted social (MWS) network. This raises a series of questions: How can one translate conclusions from partial datasets to the MWS network? More importantly, what does the MWS network look like? The first

* Corresponding author at: Asia Pacific Center for Theoretical Physics, Pohang 37673, Republic of Korea.
E-mail address: hang-hyun.jo@apctp.org (H.-H. Jo).

https://doi.org/10.1016/j.physa.2018.02.023

question has been investigated in terms of sampling biases [17–20], while the second question is largely unexplored mainly due to the lack of comprehensive datasets.

Characteristics of the MWS network are expected to be partially reflected in the empirical findings from some aspects of the network. By collecting such findings from diverse sources, we find several commonly observed features or *stylized facts* of social networks [21–23]. These include broad distributions of local network quantities [1,24], community structure [25], homophily [26,27], and intensity-topology correlations [9], etc. More recently, geographical and/or demographic information of social networks have also been explored [28–30], which are beyond the scope of this paper. One can find the previous efforts of modeling social networks: The global picture for social networks has been described by the Granovetter's hypothesis of "strength of weak ties" [31], indicating that communities of strongly connected nodes are weakly connected to each other. This picture has been empirically confirmed [9,32] and subsequently produced with computational modeling by considering cyclic and focal closure mechanisms in tie formation [22,33,34]. However, it was recently suggested that communities could be overlapping [35,36] in contrast to the picture of separate communities. This overlapping behavior is mostly due to the multilayer nature of social networks [37,38], in which each layer may correspond to a certain type of human relationship or context. This means that an individual can belong to one community in one layer but simultaneously to another community in another layer. Accordingly, dynamical models for multilayer, overlapping community structure have been introduced, while reproducing other stylized facts for local network quantities [39]. There are also several other dynamical models that partially reproduce stylized facts [40–42].

As many models mentioned above are dynamic and evolutionary in nature, they tend to take considerable amount of computational time. For relatively simpler and easier implementation, we take an alternative approach of static modeling to reproduce the stylized facts in social networks. For our model, we randomly assign a number of communities to a given set of isolated nodes using a few reasonable assumptions such that the community size is heterogeneous, and larger communities are assigned with smaller link density and smaller link weight. As we assign communities by hand rather than grow the network by means of some link formation mechanisms, our model can be called static. We also remark that the community size distribution is an input rather than an output of our model, although it is one of stylized facts. With the above mentioned few assumptions about communities, apparently realistic social network structures are generated showing most stylized facts for a wide range of the parameter space. Furthermore, thanks to the random nature of assigning communities to nodes, we can to some extent analytically calculate various local network quantities, e.g., for the assortative mixing, local clustering coefficient, and neighborhood overlap.

This static modeling approach of ours is comparable to other static modeling studies, which can be classified, but not exclusively, into four categories: (i) Erdős–Rényi (ER) random graphs, (ii) configuration models (CMs), (iii) stochastic blockmodels (SBMs), and (iv) exponential random graph models (ERGMs). The ER random graphs [43] are the simplest kind of static models, and its variants have been studied, such as graphons [44,45], weighted random graphs [46], or ER random graphs with community structure [47]. In the simplest form of CMs a binary network is constructed only by using the predetermined degree sequence of nodes, without any other correlations [48]. It has been extended for containing the arbitrary distributions of subgraphs [49], to weighted networks [50,51], or to networks with overlapping community structure [52] or with hierarchical community structure [53]. Next, the SBM was originally suggested for the community structure, characterized by a matrix consisting of the linking probabilities within communities and between communities [54–56]. As the traditional SBMs are not comparable with the empirical degree heterogeneity, the degree-corrected SBMs, by which the degree heterogeneity can be properly considered, have been studied [57,58]. The SBMs have also been extended to incorporate the overlapping communities by considering the mixed membership [59,60] or to the weighted networks, for which see Ref. [61] and references therein. Finally, the family of ERGMs has been extensively studied in social sciences [44,62] as well as in terms of statistical mechanics [63]. Here an ensemble of networks with given network features is considered according to the probability in the form of Boltzmann factor. The ERGMs have been extended for weighted networks [64] or for networks with community structure [58].

In our work, we will be exploring a different static modeling approach by explicitly considering the communities with various sizes, linking probabilities, and link weights. This way we arrive at a simple and scalable static model, which may serve as a reference system, benchmark, or testbed for further applications.

Our paper is organized as follows: In Section 2, we summarize the observed stylized facts for social networks from diverse sources. Then we introduce the community-based static model in Section 3. In Section 4, by performing large-scale numerical simulations, we find a wide range of the parameter space in which the stylized facts are reproduced. In Section 5, we present the analytical results for local network quantities. Finally, we conclude our work in Section 6.

2. Stylized facts

In Table 1 we present a summary of the commonly observed features or *stylized facts* in many digital datasets for social networks. These include the statistics of local network quantities and results for the global structure, both of which can be either topological or intensity-related.

Let us first consider topological quantities. The degree k of a node is the number of its neighbors. Degree distributions $P(k)$ in most datasets are found to be broad and overall decreasing [1,12,13,24,65]. This implies that the most probable degree or the mode of $P(k)$, denoted by m_k, is of the order of 1, leading to the fact that m_k is much smaller than the average degree $\langle k\rangle$.