SCIENCE CHINA
Information Sciences

• MOOP •

December 2016, Vol. 59 123101:1–123101:6
doi: 10.1007/s11432-015-0790-x

Structural properties and generative model of non-giant connected components in social networks

Jianwei NIU* & Lei WANG

State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191*, China*

Received May 31, 2016; accepted September 12, 2016; published online November 3, 2016

Abstract Most previous studies have mainly focused on the analyses of one entire network (graph) or the giant connected components of networks. In this paper, we investigate the disconnected components (non-giant connected component) of some real social networks, and report some interesting discoveries about structural properties of disconnected components. We study three diverse, real networks and compute the significance profile of each component. We discover some similarities in the local structure between the giant connected component and disconnected components in diverse social networks. Then we discuss how to detect network attacks based on the local structure properties of networks. Furthermore, we propose an empirical generative model called iFriends to generate networks that follow our observed patterns.

Keywords disconnected components, giant connected component, structural properties, significance profile, generative model

Citation Niu J W, Wang L. Structural properties and generative model of non-giant connected components in social networks. Sci China Inf Sci, 2016, 59(12): 123101, doi: 10.1007/s11432-015-0790-x

1 Introduction

In recent years, Social Networking Service (SNS) is developing rapidly due to its wide applications. Social networks, such as Facebook, Twitter in USA, and Renren, Sina Weibo in China, have generated a huge amount (PB level) of data which makes observations and experiments on the components in large scale possible. New observations and surprising results have been proposed, such as the power-law degree distribution [1], the densification power law and shrinking diameter [2], and the assortative or disassortative mixing patterns [3], which reveal interesting properties of social networks.

However, the above-mentioned previous studies are mainly focused on the largest connected component in a network (graph), which contains a significant portion of nodes in the network. But in real networks, there often exist many disconnected components (DCs), which are relatively independent, small components that do not connect with any other components. In earlier work, some interesting properties of DCs have been discovered, such as the star-shaped "middle regions" [4], the "gelling point" of the network diameter [5], and the final size power law [6]. In this paper, we turn our attention to the analyses of these DCs. We want to know the answers to these questions, what information can we obtain by studying

* Corresponding author (email: niujianwei@buaa.edu.cn)

Table 1 The datasets we study

Name	$\|N\|$	$\|E\|$	Type
Renren	516765	6866141	Undirected
Facebook	4039	88234	Undirected
Twitter	81306	1768149	Directed

the local structure of the DCs in a network? How can we utilize the local structural properties of DCs in some practical scenarios? Can we design a generative model to generate a network that follows the observed patterns?

The answers to these questions may help us detect abnormalities, for instance, hunting for Botnets [7], unusual groups and network attacks in social networks [8]. During the evolutionary process of networks, if some components demonstrate local structural properties significantly different from the expected cases, it is reasonable for us to suspect that such components are likely to have some abnormal properties. For example, the executives of commercial companies may get to know the anomalous behaviors of customers or groups by analyzing their online activities and the structural properties of networks [9]. Furthermore, we can utilize these patterns to reduce message and balance load in distributed graph computation [10], predict the future states of a network in some situations [11] and evaluate network communities [12].

In this paper, first, we study the structural and statistical properties of different components in some social networks. Then, we discuss how to employ the observed structural properties to detect network attacks. Finally, we propose a network generator called iFriends and conduct some empirical validations of our generator.

Our contributions mainly include the followings:

• We observe the similarities in local structure between the giant connected component (GCC) and DCs in some social networks.

• We introduce a method of detecting network attacks based on the structure properties of networks.

• We propose a generative model that can produce networks with our observed patterns and empirically validate it. Networks generated by the model, which are very difficult or impossible to be obtained from the real world, can be used to evaluate the algorithms or models of hunting for unusual groups, network attacks, etc. in social networks.

2 Datasets description

For the all three datasets that we use in our study, two of them are publicly available. The Renren dataset contains the records of the who-follows-who relations among users on Renren, which is an online real-name social networking website in China. We obtain this dataset through our collaborations on research with the Renren Corporation. The Facebook dataset describes a friendship network of the famous website Facebook. The Twitter dataset describes the social network of Twitter users and their friendship connections. The Facebook dataset and Twitter dataset can be obtained from http://snap.stanford.edu/data. The details about the datasets we use are displayed in Table 1.

3 Structural properties of components

From [13] we can know that a typical social network usually has a GCC which involves a significantly large fraction of nodes and some DCs which are defined as the small components that are not connected to any other components in the network. In this section, we mainly focus on the local structure of the GCC and DCs.

The adding process of nodes and edges usually follows a same or similar patten in the same network [14]. So we want to find if there exists any similar local structure between the GCC and DCs. In early research, there were some results on the significance profile (SP) of the graph [15]. To study the SP of the GCC and DCs, we first divide the different components of a graph into the GCC and DCs as subgraphs.

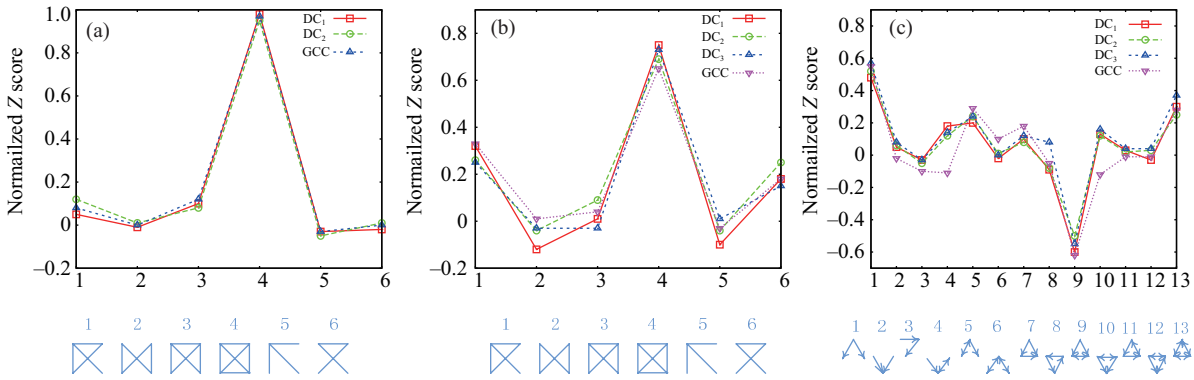

Figure 1 The SP of subgraph for different components of a graph. (a) Renren; (b) Facebook; (c) Twitter.

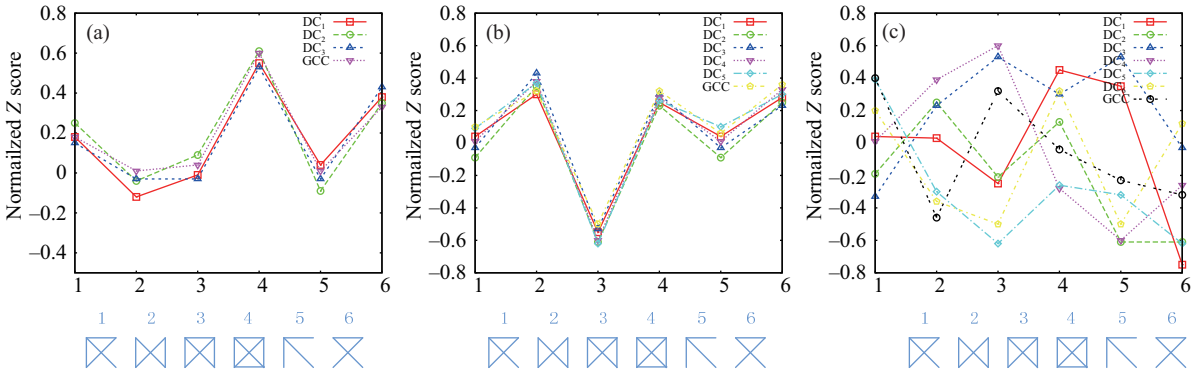

Figure 2 The SP of a network before and after a network attack. (a) Before an attack; (b) after a random-failure; (c) after a deliberate attack.

Then we calculate the SP of these components, which means the importance of 13 different triads for a directed graph and 6 different quadruples for an undirected graph. To calculate the SP of a graph, we first randomize the graph, and for each of the 13 triads for a directed graph and the 6 quadruples for an undirected graph, we compute the Z_i as follows:

$$Z_i = \frac{\text{Nreal}_i - \langle \text{Nrand}_i \rangle}{\text{std}\,\langle \text{Nrand}_i \rangle}, \tag{1}$$

where Nreal_i means the times a specific triad or quadruple appears in a graph, and Nrand_i means the times a specific triad or quadruple appears in the randomized graph. $\langle \text{Nrand}_i \rangle$ and $\text{std}\,\langle \text{Nrand}_i \rangle$ denote the average value and standard deviation of Nrand_i, respectively. Then we normalize the array Z to length 1, and get the SP_i:

$$\text{SP}_i = \frac{Z_i}{(\sum Z_i^2)^{\frac{1}{2}}}. \tag{2}$$

From Figure 1 we can see the result of these datasets. We can observe that in these datasets, the GCC and DCs share the similar SP_i. So we can draw the conclusion that in these networks there are some local structural similarities between the GCC and DCs. In our study, we also find that in some networks the triad significance profile (TSP) or quadruple significance profile (QSP) of DCs are quite different from the GCC. These networks are often the topology of "specialized" networks like collaboration networks, email networks. The underlying reason may be that these DCs are set for "especial" purpose, so they are separated from others. But in networks like topology of relationship on SNS, citation network, website links, we can always see the similar TSP or QSP between the GCC and DCs. We may explain it as the nodes in these networks are "un-organized" individuals who just make their decisions randomly, so there are not so many "special" DCs in these networks.

4 Detecting network attacks based on structure properties of networks

In this section, we discuss how to detect network attacks based on our discoveries in Section 3. In past studies, researchers studied the robustness of networks [16], and found that the network with the power-law distribution usually has better robustness in the situation of random-failures than deliberate attacks. So how will the components of a network be like after a random-failure or a deliberate attack? How can we separate deliberate attacks from random-failures in practical scenarios? To address this issue, we analyze a network as follows:

- Compute the SP of its GCC and DCs.
- Randomly make a part of the nodes of a network into "disabled" ones, meaning moving all the links connected to the selected nodes. We use this method to simulate the random-failures in the network. Then we compute the SP of its GCC and DCs after the "random-failure".
- We rank the nodes in the network according to their degrees, and make the part of nodes with relatively high degrees into "disabled" ones. We use this method to simulate the deliberate attack to the network. Then we compute the SP of its GCC and DCs after the "deliberate attack".

We choose a peer-to-peer (P2P) network and simulate the processes of a random-failure and a deliberate attack. After the simulation, we calculate the SP of the GCC and DCs under the two types of attacks. The results are displayed in Figure 2.

From Figure 2, we can see that before network attacks, there exist some local structural similarities between the GCC and DCs in the P2P network. After a random-failure or a deliberate attack, the SP of the GCC and DCs changes. However, the P2P network has similar local structures between the GCC and DCs after a random-failure, whereas the P2P network displays a significant difference in local structures between the GCC and DCs after a deliberate attack. So in a network incident, it can be possible for us to judge whether it is caused by a random-failure or a deliberate attack by analyzing the local structures of the GCC and DCs.

5 Proposed generative model

In this section, our goal is to propose an empirical generative model (we term it iFriends) to generate a large network that has the above-mentioned attributes. So the following principles should be considered: (a) the property of growth, (b) the properties of nodes and edges, and (c) preferential attachment in the adding process of new nodes.

5.1 Model description

In the iFriends model, we use a concept of social-distance to measure the similarity of two nodes by comparing their social property vectors. A social property vector is an individual property, and each node N_i is born with a random n-dimensional vector $p_i = (v_1,v_2,...,v_n)$. Each dimension in p_i has a value that represents a characteristic such as geographic location, hobbies and so on. So we define the social-distance between two nodes N_i with the social property vector $(v_1,v_2,...,v_n)$ and N_j with the social property vector $(u_1,u_2,\ldots,u_n)$ as

$$D_{i,j} = \sqrt{(v_1-u_1)^2+(v_2-u_2)^2+\ldots+(v_n-u_n)^2}. \tag{3}$$

From the definition we can know that the more similarities that nodes i and j share, the smaller value of $D_{i,j}$ is. We use this social-distance to determine the possibility for a pair of nodes to create a edge between them, i.e., the bigger value of $D_{i,j}$ is, the less possible the edge between nodes i and j will be created with.

We start with a network $G = (V, E)$, where $|V| = 1$ and $E = \emptyset$, and the newcomer nodes arrive one by one. For each node, joining a network includes two steps.

Step 1. Edge generation based on nodes' similarities. When node N_i arrives, it will create connections with some existing nodes. The possibility of the creation of the connection with N_j is $p_a \cdot f(D_{i,j})$, here

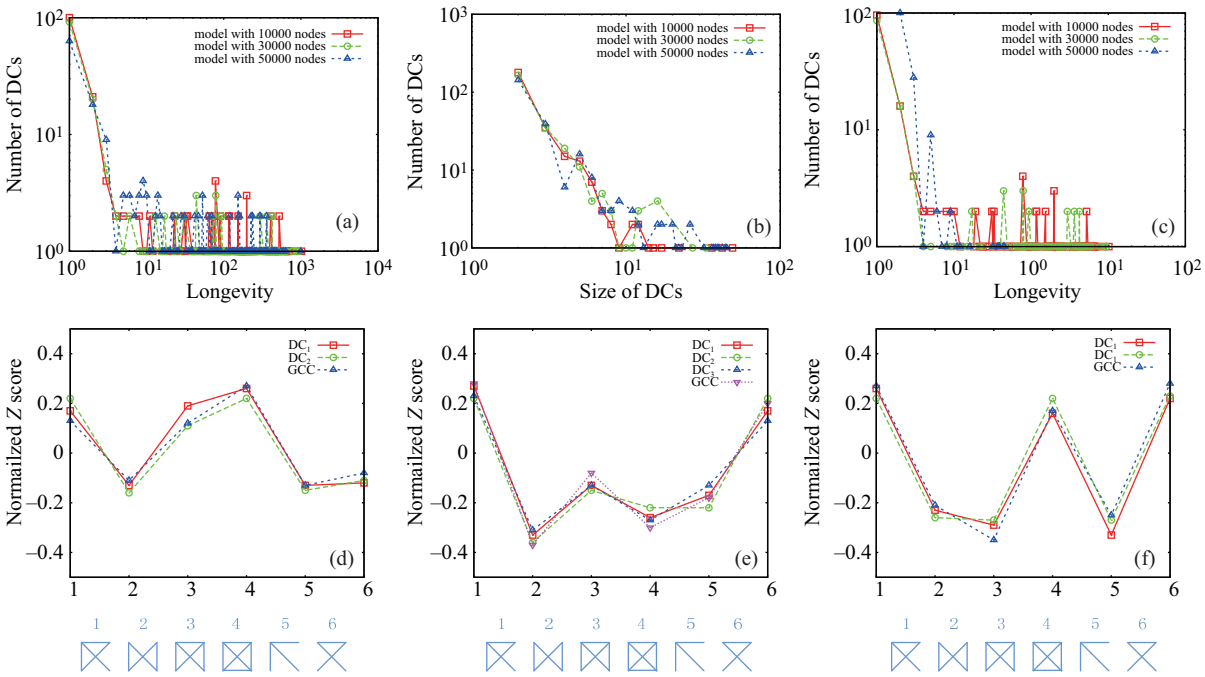

Figure 3 Results for proposed *iFriends* model. (a) Longevity distribution of DCs; (b) final size distribution of DCs; (c) size distribution of DCs merged by GCC; (d) network with 10000 nodes; (e) network with 30000 nodes; (f) network with 50000 nodes.

p_a is a empirical constant (we set $p_a = 0.8$ here). The value of $f(D_{i,j})$ is inversely proportional to the value of $D_{i,j}^2$.

Step 2. Edge generation based on mutual friends. For the newcomer node to choose from existing nodes to form edges, there will be a further influence on the graph. Let S_A be the set of all the chosen nodes in step 1, and for each $N_j \in S_A$, we traverse the set of Neighbor(N_j), i.e., the set of nodes that connect directly to N_j. Obviously, two persons who don't know each other are likely to become friends if they share a lot of friends. Based on this consideration, for each $N_k \in$ Neighbor(N_j), the possibility of the creation of the edge $e(N_i,N_k)$, $P\{e(N_i, N_k)\}$, is computed as

$$P\{e(N_i, N_k)\} \propto \frac{\text{Neighbor}(N_i) \cap \text{Neighbor}(N_k)}{\text{Neighbor}(N_i) \cup \text{Neighbor}(N_k)}. \tag{4}$$

As mentioned in the previous section, the largest connected component in a network is called GCC and the other components are called DCs. In this generative model, when a newcomer node forms edges to some existing nodes in the network, it can be connected to the GCC or existing DCs. A DC is claimed to be dead if it is connected to the GCC or merges with other DCs. In the former situation, the dead DC and the GCC become a new GCC. In the latter case, those merged DCs form a new born DC.

5.2 Empirical validation

The network generated by our proposed iFriends model should obey structural properties observed in Section 3 as well as properties observed in previous work, such as the properties observed in [6].

To evaluate the iFriends model, we generate networks that have 10000, 30000 and 50000 nodes, respectively while each node has a 3-dimensional social property vector with $p_a = 0.8$. We then compute the longevity distribution, final size distribution and size distribution of DCs merged by the GCC. Furthermore, we compute the SP of the GCC and DCs. The results are shown in Figure 3.

From Figure 3(a), we can see the apparent decaying trend with oscillations at the tail of the curve. From Figures 3(b) and (c), we can discover that the portion of merging among DCs is relatively small and the majority of DCs are absorbed by GCC. These observations are consistent with the discoveries in [6].

Next, we analyze the local structure of networks generated by our iFriends model. From Figures 3(d)–(f), it is obvious that in all three undirected networks, the GCC and DCs share the similar QSP. That is to say, there exist some local structural similarities between the GCC and DCs in these networks, which demonstrates that our generated networks have the attributes discussed in Section 3.

6 Conclusion and future work

In this paper, we made some empirical observations and analyses on structural properties of DCs in diverse social networks and discovered the similarities in the local structure between the GCC and DCs. Then we propose a method of detecting network attacks based on structural properties of DCs. Furthermore, we proposed a generative model which can fit our observations, and conducted empirical validations of it. We believe that our generative model and observations will shed light on understanding the evolution of the DCs, and could be applied in some scenarios. It is promising to employ our results to design some social network-based applications in the future.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant Nos. 61572060, 61190125, 61472024) and CERNET Innovation Project 2015 (Grant No. NGII20151004).

Conflict of interest The authors declare that they have no conflict of interest.

Supporting information The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Faloutsos M, Faloutsos P, Faloutsos C. On power-law relationships of the internet topology. SIGCOMM Comput Commun Rev, 1999, 29: 251–262
2 Leskovec J, Kleinberg J, Faloutsos C. Graphs over time: densification laws, shrinking diameters and possible explanations. In: Proceedings of the 11th International Conference on Knowledge Discovery and Data Mining, Chicago, 2015. 177–187
3 Ahn Y Y, Han S, Kwak H, et al. Analysis of topological characteristics of huge online social networking services. In: Proceedings of the 16th international conference on World Wide Web, Banff, 2007. 835–844
4 Kumar R, Novak J, Tomkins A. Structure and evolution of online social networks. In: Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining, Philadelphia, 2006. 611–617
5 McGlohon M, Akoglu L, Faloutsos C. Weighted graphs and disconnected components: patterns and a generator. In: Proceedings of the 14th International Conference on Knowledge Discovery and Data Mining, Las Vegas, 2008. 524–532
6 Niu J, Peng J, Tong C, et al. Evolution of disconnected components in social networks: patterns and a generative model. In: Proceedings of the 31st International Performance Computing and Communications Conference, Austin, 2012. 305–313
7 Yan G H. Peri-Watchdog: Hunting for hidden botnets in the periphery of online social networks. Comput Netw, 2013, 57: 540–555
8 Shrivastava N, Majumder A, Rastogi R. Mining (social) network graphs to detect random link attacks. In: Proceedings of the 24th International Conference on Data Engineering, Cancun, 2008. 486–495
9 Leskovec J, Adamic L, Huberman B. The dynamics of viral marketing. ACM Trans Web, 2007, 1: 1–39
10 Yan D, Cheng J, Lu Y, et al. Effective techniques for message reduction and load balancing in distributed graph computation. In: Proceedings of the 24th International Conference on World Wide Web, Florence, 2015. 1307–1317
11 Dong Y, Zhang J, Tang J, et al. CoupledLP: Link prediction in coupled networks. In: Proceedings of the 21th SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, 2015. 199–208
12 Yang J, Leskovec J. Defining and evaluating network communities based on ground-truth. Knowl Inf Syst, 2105, 42: 181–213
13 Broder A, Kumar R, Maghoul F, et al. Graph structure in the web. Comput Netw, 2000, 33: 309–320
14 Ugander J, Backstrom L, Marlow C, et al. Structural diversity in social contagion. Proc Natl Acade Sci, 2012, 109: 5962–5966
15 Milo R, Itzkovitz S, Kashtan N, et al. Superfamilies of evolved and designed networks. Science, 2004, 303: 1538–1542
16 Carlson J M, Doyle J. Complexity and robustness. Proc Natl Acade Sci, 2002, 99: 2538–2545