Patterns and A Generator of Social Networks: From the Perspective of Non-giant Connected Components

Jianwei Niu*, Jing Peng*, Chao Tong*

*School of Computer Science and Engineering, Beihang University, China

niujianwei@buaa.edu.cn, pj.hades@gmail.com, tongchao@buaa.edu.cn

***Abstract*—What patterns do non-giant connected components in a graph or network have? How do the non-giant connected components behave during their evolution over time? How can we model the time-evolving patterns of non-giant connected components? These questions are important for understanding the evolution of social networks, but they were seldom studied in previous work, which focused mainly on the giant connected component. In this paper, we study three real-world networks, and analyze some patterns of non-giant connected components. The main contributions of our work include the following aspects. First, we find that many non-giant connected components cannot stay in the networks for a long time. Most of them merge with one another or with the giant connected component. Second, we find that when those non-giant connected components die, the distribution of their node number follows a power law. Third, we design a graph generator to reproduce the observed patterns.**

***Keywords*—Social networks, network analysis, non-giant connected components, network evolution, graph generator**

I. INTRODUCTION

Social network analysis has become an attractive and promising field in the recent decade. Researchers have already proposed many interesting or even surprising results. How far away are we from one another in social networks? How are the nodes connected in networks? And can we reconstruct the large-scale real-world networks for further research work? People have made great effort to answer these questions, and reported various patterns and models. For instance, the small-world phenomenon [1] and its related Watts-Strogatz model [2], the assortative mixing patterns [3], the power-law degree distribution [4], the Barabási-Albert model [5], and the preferential attachment model [6].

Compared with these early studies, most recent research work mainly focus on the large-scale graph or network mining, in order to understand both static and dynamic network patterns in macroscopic and microscopic manners. Valuable conclusions have been obtained through experiments on real-world data, which is often collected from various online social network applications. For instance, the densification power law and shrinking diameter [7], the multi-scaling behavior [8], and the edge weight power law [9].

However, these studies all made explicit or implicit assumptions that the observations or experiments were conducted only on the giant connected components of the analysed networks. According to [10], real-world networks often have many non-giant connected components, which are much smaller than the giant connected component, and have no connections with it. In real-world networks, these non-giant connected components may represent different groups, such as groups of users on a website, groups of computers on the Internet, or groups of consumers or merchants in the e-marketing scenarios. Understanding the properties and patterns of such non-giant connected components may be helpful to various fields, such as electronic commerce [11], viral marketing [12], attack defense [13], [14], anomaly detection [15] and prediction of the future growth of online groups [16].

In this paper, we focus on the non-giant connected components. We study their properties by analysing the lifetime and size distributions of those components. We also propose a generator and validate it. Our major contributions are summarized as follows: 1) We find that many non-giant connected components cannot stay in networks for a long time. Most of them merge with one another or the giant connected component; 2) We find that when those non-giant connected components die, the distribution of node number in them follows a power law; 3) We design a graph generator to reproduce the observed patterns and validate it.

The rest of this paper is organized as follows. Section II lists some related work. Section III describes the three datasets we study in this paper. In Section IV, we provide some overview about the connected components in our datasets. In Section V and VI, we study the lifetime and the number of nodes of non-giant connected components during their evolution, respectively. Section VII introduces our proposed model. We conclude our work in Section VIII.

II. RELATED WORK

In this section, we review some properties or patterns together with some generators that are proposed by previous studies.

The static and dynamic properties of networks have been investigated by many researchers in this field. Typical properties include the power laws [4], small-worldness [1], densification power law and shrinking diameter [7]. For the evolution of networks, Kumar et al. [17] studied some basic time-evolving structural properties like density, diameter and edge reciprocity of Yahoo!360 and Flickr. McGlohon et al. reported in [18] that real-world networks often demonstrate a "gelling-point", which is the time at which the giant connected component is formed, and after that the networks begin to show properties like shrinking diameter. To the best of our knowledge, it seems that only [17] and [18] explicitly discussed the properties of non-giant connected components. Ugander et al. [19] found a strong relation between the number of connected components in an individual's contacts and the probability of contagion. Leicht et al. [20] analysed some citation networks with several optimization methods to characterize their structural evolution. Leskovec et al. [21] explored the microscopic structural changes that may lead to global phenomena by using the maximum likelihood estimation. People also proposed diverse network generators when exploring the formation of networks. Besides the classic Erdös-Rényi (ER) model [22] and Watts-Strogatz (WS) model [1], in the recent decade, the community-guided attachment (CGA) model and the forest fire model [7], together with the butterfly model [18] were proposed. Du et al. [9] designed a utility-based model to reproduce their observations such as the clique participation law.

However, to the best of our knowledge, the previous studies seldom discuss the evolutionary patterns of non-giant connected components. Our work differs from all the previous work in that, we focus explicitly on the evolutionary patterns of non-giant connected components, and we study these patterns especially on the dead (merged) non-giant connected components. Moreover, we propose a generator that can reproduce our observations about non-giant connected components.

III. Datasets Description

In this paper, we study three datasets that are publicly available.

A. citeulike

The *citeulike* dataset (http://www.citeulike.org/faq/data.adp) contains information about the relations between users and tags from the famous scientific reference sharing site Citeulike. The dataset records the IDs of users who create tags along with the time at which the tags are created. We process the dataset to create a network, where each node represents a user or a tag, and there is an edge between two nodes if a user creates a tag. The creation time is also recorded with each edge. The resulting network has 3,826,344 nodes (users and tags) and 4,835,698 edges (relations), and spans about 700 days.

B. imdb

The *imdb* (http://www.autonlab.org/autonweb/17433.html) dataset describes the relations between actors or actresses and the movies from the year 1888 through 2008. We process the dataset in a similar way as before: actors and actresses are represented by nodes, and if an actor or actress plays a role in a film, we create an edge between the two corresponding nodes. Then we obtain a network with 3,228,008 nodes and 8,008,134 edges.

C. renren

The *renren* dataset is crawled from the real-name social networking site Renren (http://www.renren.com/), and records the who-follows-whom relations on the website. The dataset describes the two-year growth of the Renren network from 2005 to 2007. In the Renren network, nodes represent users, and if two users are friends on the website, there is an undirected edge between them. There are 516,765 nodes and 6,866,141 edges in the network.

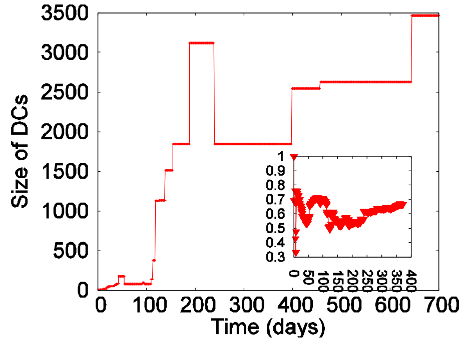

Figure 1. The growth of the second largest connected component and the node fraction of the giant connected component in the *citeulike* dataset

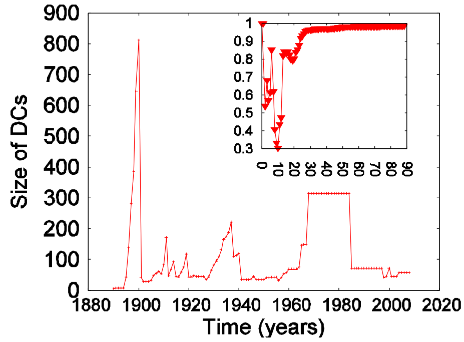

Figure 2. The growth of the second largest connected component and the node fraction of the giant connected component in the *imdb* dataset

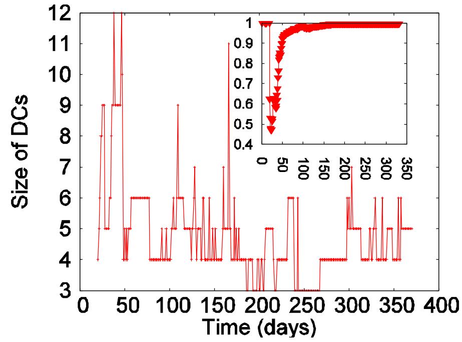

Figure 3. The growth of the second largest connected component and the node fraction of the giant connected component in the *renren* dataset

IV. Overview of Connected Components

In this section, we study some basic properties of the connected components in our datasets. We focus mainly on the formation of the giant connected component and the size evolution of the second largest non-giant connected component.

Social networks often demonstrate a single giant connected component which contains a significant portion of nodes. The giant connected component keeps adopting new nodes as the network grows, and therefore becomes larger and larger during the evolution of the network. We display the size evolution of the second largest non-giant connected component, along with the fraction of nodes in the giant connected components of our datasets in Figures 1, 2 and 3, respectively.

From Figures 1, 2 and 3, we can see apparently that the "gelling-point" [18] phenomenon, especially in *imdb* and *renren*. In the insets of the above three figures, we plot the fraction of nodes that a giant connected component contains during its growth in each of the three datasets. The *citeulike* dataset could possibly have only a short observation period, so that it is hard to find an obvious "gelling-point". However, in either of the other two datasets *imdb* and *renren*, we can see that after the "gelling-point", the size of the giant connected component begins to increase sharply, until covering over 90% of the nodes in the two networks. Because of the fast growth of the giant connected component, the size of the second larges connected component (non-giant) is limited, and cannot increase too much after the "gelling point".

Shifting our focus from the giant connected component to the non-giant ones, we begin to study the properties and patterns of the non-giant connected components in the following sections.

V. LIFETIME OF NON-GIANT CONNECTED COMPONENTS

In this section, we study the lifetime of non-giant connected components. Here we define the lifetime of a non-giant connected component as the length of the period starting from its emergence in the network until it merges with other components (giant component or others). According to [17], we know that the connected components in social networks, whether the giant one or not, often merge with one another. Then we ask the question that what patterns we can obtain from the evolution of non-giant connected components during their lifetime.

We focus mainly on the lifetime distribution of non-giant connected components. For each dataset, we start from an empty network and maintain a disjoint-set data structure [23] to record the information of non-giant connected components. Each set in the disjoint-set data structure then represents a non-giant connected component. We then follow the order that edges join the network, updating the disjoint-set data structure if the edge connects two non-giant connected components. If the edge connects a single node and a non-giant connected component, we increase the lifetime length of that component by one. If the edge connects two non-giant connected components, those two components then reach the end of their lifetime, and a new larger component is created from their combination. We set the lifetime of the new component to be one. In this way, when we go over all the edges in a dataset, we know exactly how long a non-giant connected component is during the observation period, and then we can easily compute the distribution of lifetime of these non-giant connected components. Note that the unit of the lifetime values depends on the granularity of datasets. In *citeulike* and *renren*, the edge creation time is accurate to day, but in *imdb*, that time is accurate to year.

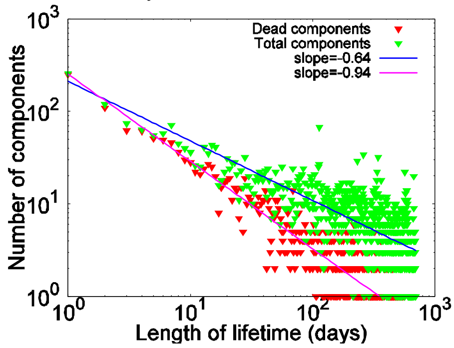

Figure 4. Lifetime distribution of non-giant connected components in the *citeulike* dataset

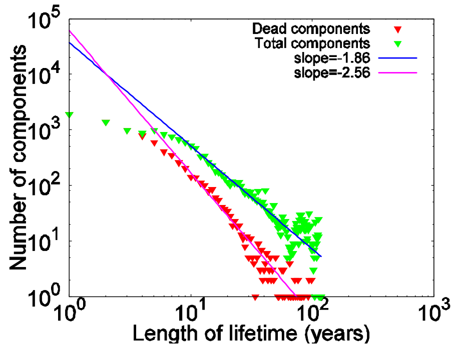

Figure 5. Lifetime distribution of non-giant connected components in the *imdb* dataset

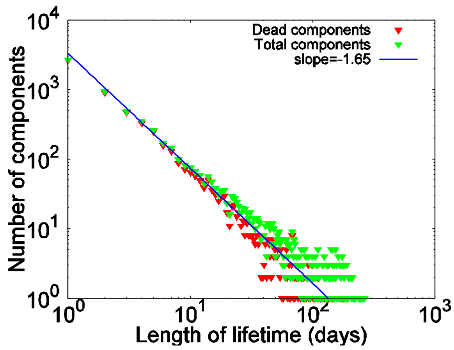

Figure 6. Lifetime distribution of non-giant connected components in the *renren* dataset

We display the lifetime distributions of *citeulike*, *imdb* and *renren* in Figures 4, 5 and 6, respectively. In each of these three figures, we plot two distribution curves, one for the lifetime distribution of all non-giant connected components, another for the lifetime distribution of only the dead components, along with the fitted lines that indicate the decaying rate. In these three figures, we can see some common patterns. All curves in these three figures are decaying apparently. This means that non-giant connected components with shorter lifetime are more in number, compared with those that have longer lifetime. From the comparison of the two distributions on each figure, we can see that the two curves are approximate to each other at first, and then diverge gradually as the lifetime length increases. This phenomenon is more obviously observed in *citeulike* and *imdb* datasets. This means that for shorter lifetime values, the portion of dead non-giant connected components are greater compared with components that have longer lifetime.

From the tails of these curves, we can infer that in all these networks, there exist some non-giant connected components that have a long lifetime length. These components may represent small groups, such as groups of users in *renren*. These non-giant connected components stay in the networks independently without forming links with other components. The only condition for them to reach a long lifetime is that the nodes inside those components create no further edges connecting nodes in other components. However, this condition can seldom hold in reality. Therefore, most of the non-giant connected components cannot reach a long lifetime during their evolution.

VI. SIZE OF NON-GIANT CONNECTED COMPONENTS

In this section, we shift our focus to the size evolution of the non-giant connected components, that is, the size of those components just before those components are about to merge with one another. The size of the non-giant connected components indicates how large they can grow during their lifetime in the networks, and its distribution can be regarded as a description of the evolution of those components.

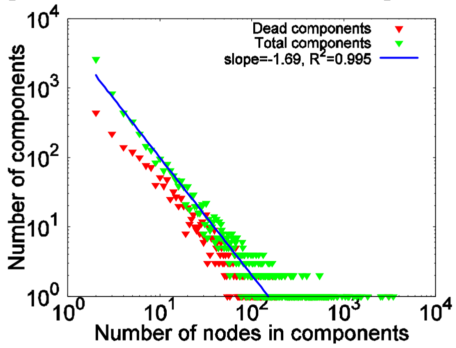

Figure 7. Size distribution of non-giant connected components in the *citeulike* dataset

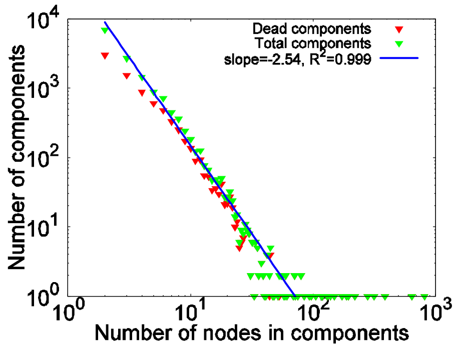

Figure 8. Size distribution of non-giant connected components in the *imdb* dataset

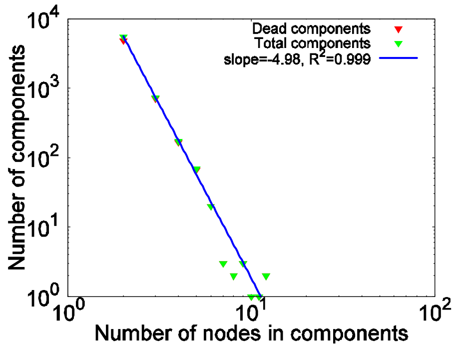

Figure 9. Size distribution of non-giant connected components in the *renren* dataset

We still use the disjoint-set data structure, and follow the same procedures as in the previous section to compute the size distributions. When two or more non-giant connected components merge (i.e., they are dead), we record their sizes. When we reach the end of the observation period of the dataset, we compute the size distribution of non-giant connected components. We show the result of our three datasets in Figures 7, 8 and 9, respectively. In each of these three figures, we show both the size distribution of all non-giant connected components, along with the size distribution of only the components that merge with one another.

There are some common patterns in Figures 7, 8 and 9. First, according to the fitted lines, we can see that the size distributions of non-giant connected components in all our datasets follow a power law, where the slopes are -1.69, -2.54 and -4.98, respectively, and the coefficients of determination are 0.995 in *citeulike*, 0.999 in both *imdb* and *renren*. The power law of size distribution shows the relation between the component size and the number of components of that size. Suppose that s_c denotes the size of a non-giant connected component, and n_c denotes the number of non-giant connected components whose sizes are s_c. Then the relation between n_c and s_c can be expressed as a power law $n_c \propto 1/s_c^{\,x}$, where the exponent x ranges from -1 to -5. Second, from the comparison between all non-giant connected components and the merged ones (the two curves on each figure), we can see that the number of merged components are very close to the total number of components. That is to say, the majority of the non-giant connected components are merged with one another, indicating that the components that remain in the networks at the end of the observation period are small in number.

As we have observed in the lifetime and size distributions, most of the non-giant connected components cannot stay in the networks for a long time, and most of them are small when they are merged. A possible explanation could be that, the larger a non-giant connected component grows, the more likely it will die out since a single connection to any node in another component will end its growth.

VII. PROPOSED MODEL

In this section, we propose our model to reproduce the aforementioned observed patterns. The generated network by our model should reveal the following properties:

- The decaying curve in the lifetime distribution of non-giant connected components;
- The power law in the size distribution of non-giant connected components;
- The proximity between the number of dead non-giant connected components and the total number of non-giant connected components.

In our model, to make the non-giant connected components merge with one another, we let some new nodes act as "bridges" between several non-giant connected components. In order to see a power law in the size distributions, we make the new nodes follow some preferential attachment rules, selecting components based on their sizes.

A. Model Description

Based on the above considerations, we introduce our model. We begin with a network which has only one node and no edge. New nodes join the network one at a time. A newcome node *v* first chooses a starting node *s* uniformly at random from all existing nodes, and begins its visit from *s*. In each step of the visit, *v* goes over all the neighbor nodes *x* of the currently visited node *u*, and decides to form an edge to *x* according to probability p_{link}, which is a parameter of our model. If *v* forms an edge to *x*, *x* will be added to the list of nodes that will be visited next. After *v* goes over all the neighbor nodes of *u*, it visits recursively all the nodes on the list. This process proceeds until no more node is chosen. When *v* finishes its visit, a *turn* of visit is done. The node *v* then decides if it will take one more turn of visit by selecting another starting node according to another model parameter p_{turn}. The pseudo-code of our model is shown in Figure 10.

B. Model Validation

We now validate our model empirically. We run our model with n=5000, p_{turn}=0.67 and p_{link}=0.3, producing a network with 5000 nodes. We follow the procedures discussed in the previous sections to compute the lifetime and size distributions of non-giant connected components in the produced network, and show the results in Figures 11 and 12. We also compute the degree distribution of the generated network in Figure 13.

From Figure 11, we can see that the curve of the lifetime distribution of non-giant connected components demonstrates a decaying pattern, indicating that short-lived non-giant connected components are more in number. From Figure 12, we can observe the power law again. In these two figures, the numbers of merged (dead) components are close to the total number of components, indicating that most of the non-giant connected components are merged with one another during the evolution of the network. Figure 13 shows the degree distribution of the generated network. From Figure 13 and the fitted line, we see that the degree distribution of the generated network follows approximately a power law.

VIII. CONCLUSIONS

In this paper, we investigated three real-world networks, and analysed the properties of non-giant connected components. We focused on the lifetime and size distributions of such components, i.e., how long they can stay in the networks and how large they can grow before they merge with one another. We found that many non-giant connected components can only stay for a short time in the networks, and they often contain only a few nodes when they finally merge with one another. All the three studied networks have a giant connected component, and a significant portion of non-giant connected components are merged and cannot live till the end of the observation period. We proposed a generative model which can reproduce our observed patterns. We believe that our work in this paper could be helpful for understanding the evolution of the non-giant connected components in social networks and further studies.

```
// Finish the visit of node current
// starting at node begin
def visit(G, v, s):
    // The queue of to-be-visited nodes
    frontier = Queue()
    frontier.add(s)
    // The set of visited nodes
    visited = Set()
    visited.add(s)
     G.add_edge(v, s)

    while frontier is not empty:
        // Pick a node from the
        // to-be-visited nodes
        u = frontier.get()
        // Choose from its
        // non-visited neighbors
        for not visited x in u.neighbors:
            if random() <= p_link:
                G.add_edge(v, x)
                // The node x will be
                // visited next
                visited.add(x)
                frontier.put(x)

// Generate a network with n nodes
def model(n):
    G = Graph()
     // We begin with a network
     // with only one node
    G.add_node(1)
    for v in 2 ... n:
        // Get all existing nodes
        exist = G.nodes()
        // Add the newcome node
        G.add_node(v)

        while random() <= p_turn:
            // Choose a starting node
            // uniformly at random
            s = choose(exist)
            visit(G, v, s)

    return G
```

Figure 10. The pseudo-code of our model

ACKNOWLEDGMENT

This work was supported by the Research Fund of the State Key Laboratory of Software Development Environment under Grant No. BUAA SKLSDE-2012ZX-17, the National Natural Science Foundation of China under Grant No. 61170296 and 61190120, the Program for New Century Excellent Talents in University under Grant No. NECT-09-0028.

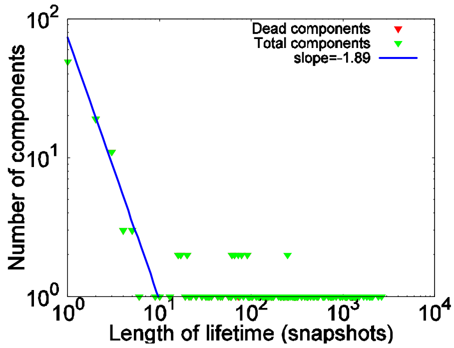

Figure 11. Lifetime distribution of non-giant connected components in the generated network

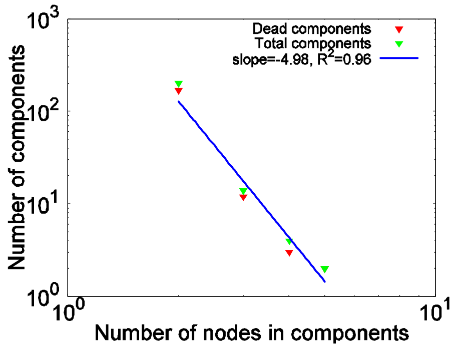

Figure 12. Size distribution of non-giant connected components in the generated network

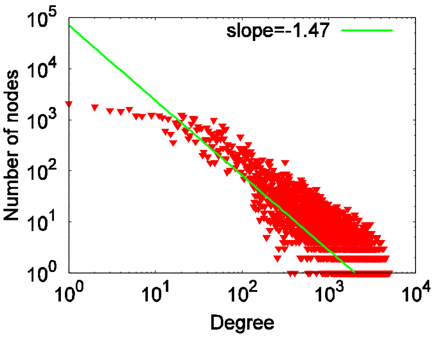

Figure 13. Degree distribution of the generated network

REFERENCES

[1] D. Watts and S. Strogatz, "Collective dynamics of small-world networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[2] A. Barrat and M. Weigt, "On the properties of small-world network models," *The European Physical Journal B–Condensed Matter and Complex Systems*, vol. 13, no. 3, pp. 547–560, 2000.

[3] M. Newman, "Mixing patterns in networks," *Physical Review E*, vol. 67, no. 2, p. 026126, 2003.

[4] M. Faloutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," *SIGCOMM Computer Communication Review*, vol. 29, no. 4, ACM, pp. 251–262, 1999.

[5] A. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[6] R. Albert and A. Barabási, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, no. 1, p. 47, 2002.

[7] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: densification laws, shrinking diameters and possible explanations," *Proceedings of the 11th International Conference on Knowledge Discovery and Data Mining (SIGKDD 05)*, ACM, pp. 177–187, 2005.

[8] Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, "Analysis of topological characteristics of huge online social networking services," *Proceedings of the 16th International Conference on World Wide Web (WWW 07)*, ACM, pp. 835–844, 2007.

[9] N. Du, C. Faloutsos, B. Wang, and L. Akoglu, "Large human communication networks: patterns and a utility-driven generator," *Proceedings of the 15th International Conference on Knowledge Discovery and Data Mining (SIGKDD 09)*, ACM, pp. 269–278, 2009.

[10] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, "Graph structure in the web," *Computer Networks*, vol. 33, no. 1, pp. 309–320, 2000.

[11] S. Guo, M. Wang, and J. Leskovec, "The role of social networks in online shopping: information passing, price of trust, and consumer choice," *Proceedings of the 12th Conference on Electronic Commerce (EC 11)*, ACM, pp. 157–166, 2011.

[12] J. Leskovec, L. Adamic, and B. Huberman, "The dynamics of viral marketing," *Transactions on the Web (TWEB)*, ACM, vol. 1, no. 1, p. 5, 2007.

[13] B. Viswanath, A. Post, K. Gummadi, and A. Mislove, "An analysis of social network-based sybil defenses," *SIGCOMM Computer Communication Review*, vol. 40, no. 4, ACM, pp. 363–374, 2010.

[14] N. Shrivastava, A. Majumder, and R. Rastogi, "Mining (social) network graphs to detect random link attacks," *Proceedings of the 24th International Conference on Data Engineering (ICDE 08)*, IEEE Computer Society, pp. 486–495, 2008.

[15] A. Lazarevic, L. Ertöz, V. Kumar, A. Ozgur, and J. Srivastava, "A comparative study of anomaly detection schemes in network intrusion detection," *Proceedings of the 3rd International Conference on Data Mining (ICDM 03)*, SIAM, 2003.

[16] S. Kairam, D. Wang, and J. Leskovec, "The life and death of online groups: predicting group growth and longevity," *Proceedings of the 5th International Conference on Web Search and Data Mining (WSDM 12)*, ACM, pp. 673–682, 2012.

[17] R. Kumar, J. Novak, and A. Tomkins, "Structure and evolution of online social networks," *Proceedings of the 12th International Conference on Knowledge Discovery and Data Mining (SIGKDD 06)*, ACM, pp. 611–617, 2006.

[18] M. McGlohon, L. Akoglu, and C. Faloutsos, "Weighted graphs and disconnected components: patterns and a generator," *Proceedings of the 14th International Conference on Knowledge Discovery and Data Mining (SIGKDD 08)*, ACM, pp. 524–532, 2008.

[19] J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg, "Structural diversity in social contagion," *Proceedings of the National Academy of Sciences*, vol. 109, no. 16, pp. 5962–5966, 2012.

[20] E. Leicht, G. Clarkson, K. Shedden, and M. Newman, "Large-scale structure of time evolving citation networks," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 59, no. 1, pp. 75–83, 2007.

[21] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins, "Microscopic evolution of social networks," *Proceedings of the 14th International Conference on Knowledge Discovery and Data Mining (SIGKDD 08)*, ACM, pp. 462–470, 2008.

[22] P. Erdös and A. Rényi, "On the evolution of random graphs," *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, vol. 5, pp. 17–61, 1960.

[23] B. A. Galler and M. J. Fisher, "An improved equivalence algorithm," *Communications of the ACM*, vol. 7, no. 5, pp. 301–303, ACM, 1964.