Coexistence of Undirected and Directed Networks in Goodreads Online Community

Jignying (Jane) Bi[1]

ABSTRACT

Goodreads is a large-scale online network. Its following-or-friend setting allows the coexistence of undirected and directed networks. A mixing-network model is built to discuss the efficiency of such network and which network structure tends to appear. The data of Goodreads' members' demographical information, list of friends and list of followings has been collected. It covers over 5 million Goodreads members, equivalent to around 8% of total 65 million population. Empirical data shows that the Goodreads network also follows the power-law distribution.

Keywords: online social network, undirected and directed network, power-law distribution.

1. INTRODUCTION

Nowadays, online social network plays an important role in information dissemination and communication. Goodreads, an online reading network, allows people to mark their reading process, exchange reading comments, and follow others' book reviews. To follow others' reviews, one could either become their friends or follow them, but not both. The former gives a bilateral relationship while the latter gives a unilateral one. (Goodreads, 2018) This setting allows the coexistence of undirected and directed networks which is this paper studying on. A mixing-network model with both types of networks is built.

Meanwhile, there are abundant empirical literatures documenting the structural features of large-scale online networks, such as Twitter, Flickr, LiveJournal, and YouTube. All these online networks share some common features, such as power-law distribution and small world-property. Similar to these networks, Goodreads also obey the common rules.

The paper is structured as follows. Section 2 gives the literature review. Section 3 builds up the mixing-network model. Section 4 talks about the empirical observations of Goodreads' topological characteristics. Conclusion is drawn in Section 5.

2. LITERATURE REVIEW

This paper is related to a number of theoretical and empirical literatures. Jackson and Wolinsky (1996) model the bilateral connection whose formation requires the consent of both parties. They study the stability and efficiency of the undirected networks and give us the predictions on which network structures are likely to form. (Matthew O. Jackson, 1996) Bala and Goyal (2000) model the unilateral connection where only the initiator of the connection bears the cost of forming and maintaining it. (Venkatesh Bala, 2000) My paper differs from these two as I am studying a community where the undirected and directed networks coexist. Specifically, members of Goodreads could either make friends with others (i.e. bilateral relationship) or follow others (i.e. unilateral relationship), but not both.

[1] UCID: 12174556; Email: jingyingb@uchicago.edu

There are also many empirical papers studying the network properties of online social networks. Alan Mislove *et al* (2007) discover and compare four online social networks, namely Flickr, YouTube, LiveJournal and Orkut. Their results confirm the power-law and small world property of the online social networks (Cha et al, 2009; Kwak et al, 2010). Bakshy, Mason, Hofman, Watts (2011) discover the power-law in Twitter's networks (Newman, 2003; Ugander et al, 2011). In addition, they also find that a tweet tends to be spread wider if the source node has more significant past influence and a larger number of followers. Hence another contribution of my paper is to discover the topological characteristics of Goodreads online network and compare with these existing empirical findings.

3. MODEL: *A Mixing-network*

Let $\mathcal{N} = \{1, \dots, n\}$ be the finite set of players. Let ij represents the link between players i and j. Meanwhile, denote g as the set where if $ij \epsilon g$, then there exists a link between i and j, either undirected or directed. G denotes the whole network. There are two possible natural states $\Phi = \{1, 2\}$, representing the types of players. Player 1 (i.e. $\Phi = 1$) is the member who initiates the connection by either following Player 2's book reviews or sending a friend request to her. If Player 1 merely chooses to follow Player 2's book reviews, a connection in the unilateral network is formed. Player 1 will benefit from accessing Player 2's updates and reviews, b. Meanwhile, he is the only person who will bear the cost to maintain the relationship, as assumed in Bala and Goyal (2000) model.

If Player 1 chooses to send a friend request, he will first suffer a sunk cost ε. This is because Player 2 always sets a question for those who want to make friends with her. For instance, *"Have we interacted before? And what is your fave quote?" or "Are you over 18? Please only proceed if you are over 18. The books I read have adult themes and language. Thank you :-)"*. Answering the question takes time and even if the request is sent with the answer, Player 2 may not accept the request, which incurs extra cost on Player 1 as ε.

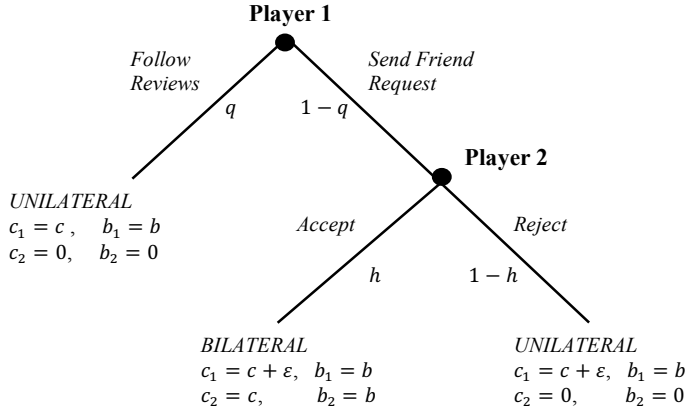


Figure 1. Strategic network formation.

Player 2 is the member who is either followed by another member or receiving Player 1's friend request. If she received the request, she will make decision to either *accept* or *reject* it. Accepting the request ends up with a bilateral connection in the undirected network, whereas rejection leads to a unilateral connection in the directed network. Should a bilateral connection

be formed, both parties will benefit b from the relationship and are willing to take the costs to maintain it. It is worth of mentioning that Player 1 starts to follow Player 2 once he sends the friend request. Whereas their final relationship (either bilateral or unilateral) depends on Player 2's response.

Assume that the probability of being Player 1 is a. The probability for Player 1 to choose to follow Player 2's reviews is q. The probability for Player 2 to accept the friend request is h. The strategic tree and relevant information is summarized in Figure 1.

Hence the expected benefits and costs for Player 1 and Player 2 are as follow.

$$E(b_1|\ Player\ 1) = qb + (1-q)hb + (1-q)(1-h)b = b \quad (1)$$
$$E(c_1|\ Player\ 1) = qc + (1-q)h(c+\varepsilon) + (1-q)(1-h)(c+\varepsilon) = c + (1-q)\varepsilon \quad (2)$$
$$E(b_2|\ Player\ 2) = q0 + (1-q)hb + (1-q)(1-h)0 = (1-q)hb \quad (3)$$
$$E(c_2|\ player\ 2) = q0 + (1-q)hc + (1-q)(1-h)0 = (1-q)hc \quad (4)$$

Therefore, the expected benefit and cost for a player are as follow.

$$E(b) = (a + (1-a)(1-q)h)b \quad (5)$$
$$E(c) = (a + (1-a)(1-q)h)c + a(1-q)\varepsilon \quad (6)$$

The utility of player i is $u_i(g) = \sum_{j:ij\epsilon g} E(b)^{t_{ij}} - E(c)d_i(G)$.
For simplification, let $m \equiv (a + (1-a)(1-q)h)$ and assume $\varepsilon \equiv m/(1-q)\,a$. By substituting (5) and (6) into the utility function, $u_i(g) = \sum_{j:ij\epsilon g} b^{t_{ij}}\, m^{t_{ij}} - m(c+1)$ (7)

PROPOSITION: In this model where the undirected and directed network coexist with each other, the unique efficient network is

(i) the complete graph if $c < b - 1 - b^2m$

(ii) a star graph if $b - 1 - b^2m < c < b - 1 + \frac{n-2}{2}b^2m$

(iii) no link if $c > b - 1 + \frac{n-2}{2}b^2m$

Proof see Appendix!

Notice that h is the probability that Player 2 will accept Player 1's friend request and subsequently a connection in the undirected network is formed. Therefore, if we could estimate the benefit, costs, q and observe the network structure, we could backward deduct the value of h. That predicts the proportion of undirected network versus the directed network. The empirical observation tells that the undirected network is far more extended than the directed one, which indicates that h has to be more than 0.5.

4. DATA COLLECTION AND ANALYSIS

Data Collection

Data was collected from the public Goodreads accounts via web scraping. I started from the user "18492568-vina" by collecting her demographical information, list of friends and list of followings (i.e. people that vina follows). The demographical information includes name, number of book ratings, number of books already read, currently reading and to read, the joined reading groups. The two lists are marked as hop 1 data as they are vina's neighborhood. Next, I collected the demographical information, list of friends, and list of followings of each hop 1 user. These new information are hop 2 data as they are 2 hops away from the starting node "18492568-vina". Then I repeated this snow-balling pattern and only completed 17500 out of

881,291 members in hop 3 due to time limitation. Table 1 summarizes the dataset and its structure.

I signed up 7 Goodreads accounts to avoid overloading the server while requesting for the data. I used 30 machines with different IP addresses to speed up the data collection. The whole collection process took 13 days from Feb 15th to March Feb 27th. Goodreads has around 65 million members in total. Due to time limitation, my data only covers 5,075,340 members in the undirected network and 159,279 members in the directed network. The procedure of data collection is summarized in Table 1.

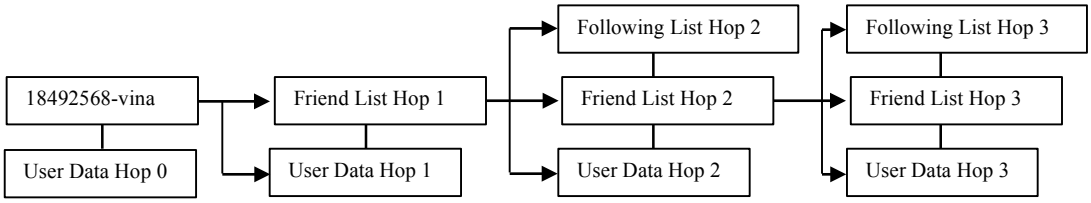
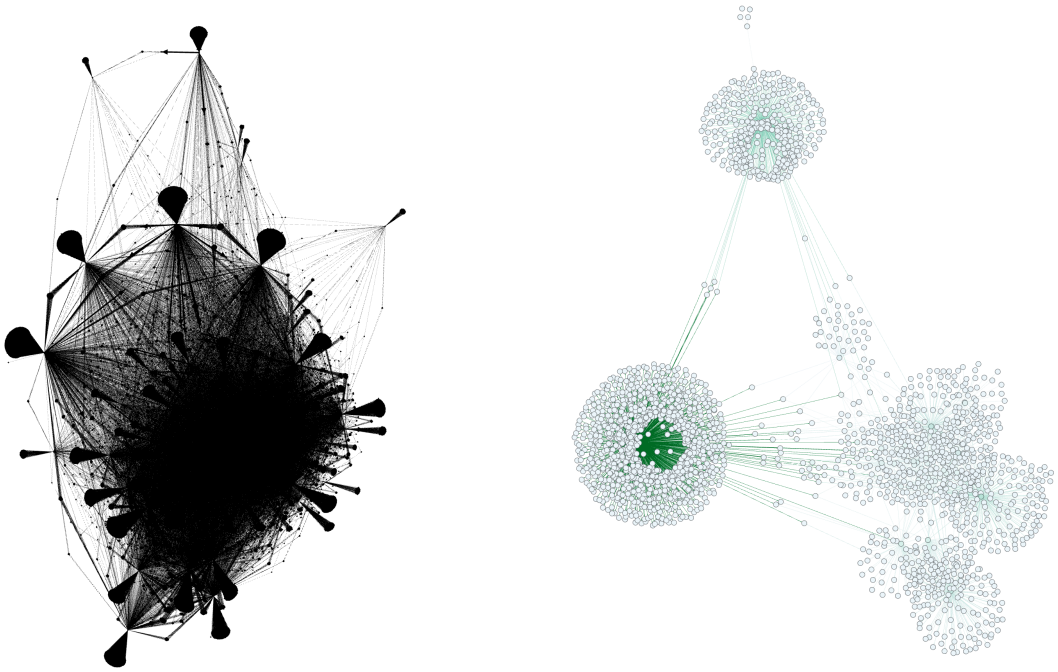


Table 1. Empirical Data from Goodreads

Data Analysis

This section provides an in-depth understanding of the topological properties of Goodreads online social network. Two key and common features of the large-scale online networks are the power-law distribution of degrees and the small-world property. The Goodreads subnetwork also follow these two rules, although my data only covers around 8% of total Goodreads population.



(a) Undirected Network (b) Directed Network

Figure 2. Sub-network Graph. In panel (a), the very dense fan-shape is where the authors attracting large number of followers. Panel (b) shows the sub-network graph of the directed network

OBSERVATION: Power-law distribution is followed in both undirected and directed networks.

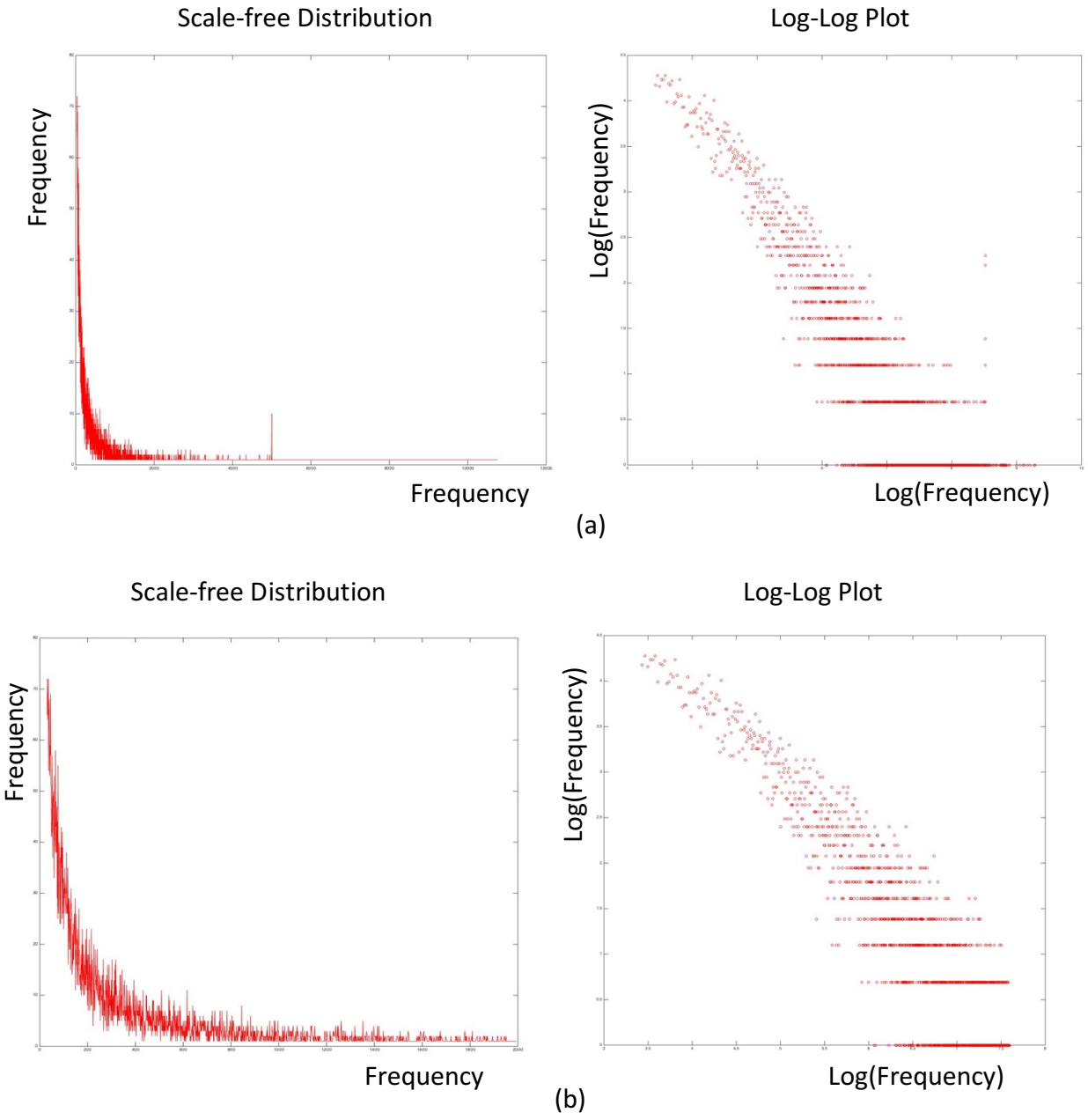


Figure 3. Power-law distribution. Panel (a) displays the scale-free distribution and the log-log plot of the undirected network degree distribution. Panel (b) displays the scale-free distribution and the log-log plot of the directed network degree distribution.

Figure 3 displays the distribution of degrees in both undirected (panel (a)) and directed (panel (b)) networks. In panel (a), the left side plot is the scale-free distribution. It displays a long and fat tail, with the shape of a power law distribution $p(x) = Cx^{-\alpha}$. The right side plot is the log-log plot, showing a roughly downward sloping linear line. This confirms that the undirected network degree distribution follows a power-law distribution. Panel (b) displays the same information for the directed network.

5. CONCLUSION

This paper explores the coexistence of undirected and directed networks. A mixing-network is built to study the efficiency of such network, and predicts which type of structures will be more likely to appear. The empirical data collected from the Goodreads confirms that this network also follows the power-law distribution.

Due to the time limitation, there are several weakness of this paper. For instance, the data only covers 8% of the total Goodreads members, which may cause biased result. Secondly, due to the data collection procedure, I'm unable to analyze the shortest paths and average paths among members because all data are within 3 hops from 18492568-vina. So, the small-world property cannot be testified here.

Future work could focus on the following three steps: (1) collect the full Goodreads dataset and figure out how to deal with large scale network dataset, as the calculation speed now is already very slow; (2) re-analyze the results above with the full dataset; (3) collect the weekly-updated "best reader" and "best reviewer" list to study he preferential attachment network evolution.

APPENDIX:

Proof of PROPOSITION.

Case I: when $c < b - 1 - b^2m$,

When there is no link between *i* and *j*, the utility of *i* is $u_i' \leq b^2m^2$. If *i* chooses to form a link with *j*, his utility is $u_i \leq bm - mc - m$. The net change in the utility after forming a link is $u_i - u_i' \geq (bm - mc - m) - b^2m^2$. Given $c < b - 1 - b^2m$, $u_i - u_i' > 0$. Therefore, all players choose to form a link with all the rest of the players in order to maximize their total utility. When this happens, the social welfare is maximized as well. Hence, the component graph is the most efficient structure.

Case II: when $b - 1 - b^2m < c < b - 1 + \frac{n-2}{2}b^2m$,

Suppose *k* direct-link are present in component, where $k > n - 1$. Then the total costs in this component is $2k(c + 1)m$. The benefit from the neighboring connections is $2kbm$, and at most $2b^2m^2(\frac{n(n-1)}{2} - k)$. Hence, the aggregate utility is upper bounded by

$2b^2m^2\left(\frac{n(n-1)}{2} - k\right) + 2kbm - 2k(c + 1)m = b^2m^2n(n - 1) - 2mk[c - (b - 1 - b^2m)]$.

Given that $c - (b - 1 - b^2m) > 0$, to maximize this aggregate utility is equivalent to maximize the value of *k*, where a "star" appears.

Case III: when $c > b - 1 + \frac{n-2}{2}b^2m$

Whenever a link is formed, the aggregate utility of two parties involved in the link is upper bounded by $2b - 2(c + 1) + (n - 2)b^2m$. Given $c > b - 1 + \frac{n-2}{2}b^2m$, this upper bound is negative. Therefore, no player is willing to connect with others. Hence, there is no link forms.

■

Bibliography

Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi (2007). Measurement and Analysis of Online Social Networks. *Proc. of IMC*.

Eytan Bakshy, Winter A. Mason, Jake M. Hofman, Duncan J. Watts (2011). Everyone's an Influencer: Quantifying Influence on Twitter. *WSDM*. Hong Kong.

Goodreads. (2018). *Help Topic*. Retrieved from Goodreads: https://www.goodreads.com/help/show/81-what-is-the-difference-between-being-someone-s-friend-and-following-thei

Haewoon Kwak, C. L. (2010). What is Twitter, a Social Network or a News Media? *International World Wide Web Conference Committee (IW3C2)*.

Haewoon Kwak, C. L. (2010). What is Twitter, a Social Network or a News Media? *World Wide Web Conference Committee*.

Johan Ugander, B. K. (2011, November 18). *The Anatomy of the Facebook Social Graph.* Retrieved from https://arxiv.org/abs/1111.4503

Matthew O. Jackson, A. W. (1996). A Strategic Model of Social and Economic Networks. *Journal of Economic Theory*, 44-74.

Meeyoung Cha, Alan Mislove, Krishna P. Gummadi (2009). A Measurement-driven Analysis of Information Propagation in the Flickr Social Network. *International World Wide Web Conference Com- mittee (IW3C2).* Madrid, Spain.

Newman, M. E. (2003, Mar 25). *The structure and function of complex networks.* Retrieved from https://arxiv.org/abs/cond-mat/0303516v1

Venkatesh Bala, S. G. (2000). A Noncooperative Model of Network Formation. *Econometrica*, 1181-1229.