Network Analysis of Three Twitter Functions: Favorite, Follow and Mention

Shoko Kato[1], Akihiro Koide[1], Takayasu Fushimi[1],
Kazumi Saito[1], and Hiroshi Motoda[2]

[1] School of Management and Information, University of Shizuoka,
52-1 Yada, Suruga-ku, Shizuoka 422-8526, Japan,
{b09032,j11103,j11507,k-saito}@u-shizuoka-ken.ac.jp
[2] Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan
motoda@ar.sanken.osaka-u.ac.jp

Abstract. We analyzed three functions of Twitter (Favorite, Follow and Mention) from network structural point of view. These three functions are characterized by difference and similarity in various measures defined in directed graphs. Favorite function can be viewed by three different graph representations: a simple graph, a multigraph and a bipartite graph, Follow function by one graph representation: a simple graph, and Mention function by two graph representations: a simple graph and a multigraph. We created these graphs from three real world twitter data and found salient features characterizing these functions. Major findings are a very large connected component for Favorite and Follow functions, scale-free property in degree distribution and predominant mutual links in certain network motifs for all three functions, freaks in Gini coefficient and two clusters of popular users for Favorites function, and a structure difference in high degree nodes between Favorite and Mention functions characterizing that Favorite operation is much easier than Mention operation. These finding will be useful in building a preference model of Twitter users.

1 Introduction

Grasping and controlling preference, tendency, or trend of the consuming public is one of the important factors to achieve economic success. Accordingly, it is vital to collect relevant data, analyze them and model user preference. However, quantifying preference is very difficult to achieve and finding useful measures from the network structure is crucial. The final goal of this work is to find such measures, characterize their relations and build a reliable user preference model based on these measures from the available data. As the very first step, we focus on Twitter data and analyze the user behavior of three functions (Favorite, Follow and Mention) of Twitter [1] from the network structural point of view, i.e., by using various measures that have been known to be useful in the graph

[1] http://twitter.com/

theory and identifying characteristic features (difference and similarity) of these measures for these functions.

User behavior of these three functions are represented by different directed graphs. Favorite function can be viewed by three different graph representations: a simple graph, i.e., single edge from a Favorer to a Favoree, a multigraph, i.e., multiple edges from a Favorer to a Favoree, and a bipartite graph, i.e., single edge from a Favorer to a Favoree treating a user with both a Favorer and a Favoree as two separate nodes. Likewise, Follow function can be viewed by one graph representation: a simple graph, i.e., single edge from a Follower and a Followee, and Mention function can be viewed by two different graphs: a simple graph, i.e. single edge from a Mentioner (sender) to a Mentionee (receiver) and a multigraph, i.e. multiple edges from a Mentioner to a Mentionee. We have created these networks from three different Twitter logs (called "Favorites network", "Followers network", and "Mentions network") and used several different measures, e.g. in-degree, out-degree, multiplicity, Gini coefficient, etc. Extensive experiments were performed and several salient features were found. Major findings are that 1) Favorites and Followers networks have a very large connected component but Mentions network is not, 2) all the three networks (both simple and multiple) have the scale-free property in degree distribution, 3) all three networks (simple) have predominant three-node motifs having mutual links, 4) Favorites network have freaks in Gini coefficient (one of the measures), 5) Favorites network have two clusters of popular users, and 6) Favorites and Mentions networks differ in structure for high degree nodes reflecting that Favorite operation is much easier than Mentions operation. In this paper, we propose to analyze multigraphs by using two new measures, i.e., correlation between degree and average multiplicity, and correlation between degree and Gini coefficient. In our experiments, we show that these measures contribute to clarify a structure difference between Favorites and Mentions networks.

Twitter, a microblogging service, has attracted a great deal of attention and various properties have already been obtained [3] [4], but to our knowledge, there have been no work to analyze the user behavior from network structural point of view. We believe that the work along this line will be useful in understanding the user behavior and helps building a preference model of Twitter users.

The paper is organized as follows. We briefly explain the various measures we adopted in our analysis in 2, three networks (Favorite, Follow, and Mention) in 3. Then we report the experimental results in 4 and provide some discussions regarding our observations in 5. We end this paper by summarizing the major finding and mentioning the future work in 6.

2 Analysis Methods

According to [1], we define the structure of a network as a graph. A graph $G = (V, E)$ consists of a set V of nodes (vertices) and a set E of links (edges) that connect pairs of nodes. Note that in our Favorites, Followers or Mentions network, a node corresponds to a Twitter user, and a link corresponds to favoring, following, or mentioning between a pair of users. If two nodes are connected by a link, they are adjacent and we call them neighbors. In directed graphs, each directed link has an origin (source) and a destination (target). A link with origin $u \in V$ and destination $v \in V$ is represented by an ordered pair (u, v). A directed graph $G = (V, E)$ is called a bipartite graph, if V is divided into to two parts, V_x and V_y, where $V = V_x \cup V_y$, $V_x \cap V_y = \emptyset$, and $E \subset \{(u, v); u \in V_x, v \in V_y$. In directed graphs, we may allow the link set E to contain the same link several times, i.e., E can be a multiset. If a link occurs several times in E, the copies of that link are called parallel links. Graphs with parallel links are also called multigraphs. A graph is called simple, if each of its links is contained in E only once, i.e., if the graph does not have parallel links. In what follows, we describe our analysis methods for each type of graphs.

2.1 Methods for Simple Graph

A graph $G' = (V', E')$ is a subgraph of the graph $G = (V, E)$ if $V' \in V$ and $E' \in E$. It is an induced subgraph if E' contains all links $e \in E$ that connect nodes in V'. A directed graph $G = (V, E)$ is strongly connected if there is a directed path from every node to every other node. A strongly connected component of a directed graph G is an induced subgraph that is strongly connected and maximal. A bidirected graph $\tilde{G} = (V, \tilde{E})$ is constructed from a directed graph $G = (V, E)$ by adding counterparts of the unidirected links, i.e., $\tilde{E} = E \cup \{(v, u); (u, v) \in E\}$. A weakly connected component of a directed graph G is an induced subgraph from V' obtained as a strongly connected component of the bidirected graph $\tilde{G}$. We analyze the structure of our networks in terms of the connectivity using these notions.

In a directed graph $G = (V, E)$, the out-degree of $v \in V$, denoted by $d^+(v)$, is the number of links in E that have origin v. The in-degree of $v \in V$, denoted by $d^-(v)$, is the number of links with destination v. The average degree d is calculated by

$$d = \frac{1}{|V|} \sum_{v \in V} d^-(v) = \frac{1}{|V|} \sum_{v \in V} d^+(v) = \frac{|E|}{|V|}. \tag{1}$$

Here $|\cdot|$ stands for the number of elements for a given set. The correlation between in- and out-degree, denoted by c, is calculated by

$$c = \frac{\sum_{v \in V} (d^-(v) - d)(d^+(v) - d)}{\sqrt{\sum_{v \in V} (d^-(v) - d)^2} \sqrt{\sum_{v \in V} (d^+(v) - d)^2}}. \tag{2}$$

On the other hand, the in-degree distribution $id(k)$ and the out-degree distribution $od(k)$ with respect to degree k are respectively defined by

$$id(k) = |\{v \in V; d^-(v) = k\}|, \quad od(k) = |\{v \in V; d^+(v) = k\}|. \tag{3}$$

We analyze the statistical properties of these degree distributions.

Network motifs are defined as patterns of interconnections occurring in graphs at numbers that are significantly higher than those in randomized graphs. In our

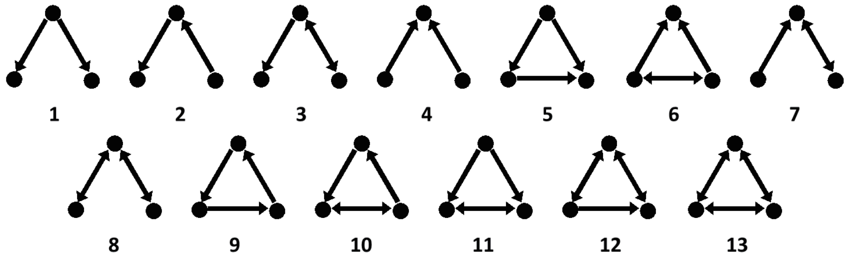

Fig. 1: Network motifs patterns

analysis, we focus on three-node motifs patterns and Figure 1 shows all thirteen types of three-node connected subgraphs (motifs patterns). According to [5], we also use randomized graphs, each node of which has the same in-degree and out-degree as the corresponding node has in the real network [6]. A significance level of each motifs pattern i is evaluated by its z-score z_i, i.e.,

$$z_i = \frac{f_i - J^{-1}\sum_{j=1}^{J} g_{j,i}}{\sqrt{J^{-1}\sum_{j=1}^{J}(f_i - J^{-1}\sum_{j=1}^{J} g_{j,i})^2}}, \tag{4}$$

where J is the number of randomized graphs used for evaluation, and f_i and $g_{j,i}$ denote the numbers of occurrences of motifs pattern i in the real graph and the j-th randomized graph, respectively. By this motifs analysis, we attempt to uncover the basic building blocks of our networks.

2.2 Visualization of Bipartite Graph

A bipartite graph is a graph whose nodes can be divided into two disjoint sets V_x and V_y such that every links connects a vertex in V_x to one in V_y. We can construct a bipartite graph from a directed graph by setting $V_x = \{u; (u,v) \in E\}$ and $V_y = \{v; (u,v) \in E\}$, and regarding that any element in V_x is different from any element in V_y. Further, according to [2], we describe a bipartite graph visualization method for our analysis. For the sake of technical convenience, each set of the nodes, V_x and V_y, is identified by two different series of positive integers, i.e., $V_x = \{1, \cdots, m, \cdots, M\}$ and $V_y = \{1, \cdots, n, \cdots, N\}$. Here M and N are the numbers of the nodes in V_x and V_y , i.e., $|V_x| = M$ and $|V_y| = N$, respectively. Then, the $M \times N$ adjacency matrix $\mathbf{A} = \{a_{m,n}\}$ is defined by setting $a_{m,n} = 1$ if $(m,n) \in E$; $a_{m,n} = 0$ otherwise. The L-dimensional embedding position vectors are denoted by $\mathbf{x}_m$ for the node $m \in V_x$ and $\mathbf{y}_n$ for the node $n \in V_y$. Then we can construct $M \times L$ and $N \times L$ matrices consisting of these position vectors, i.e., $\mathbf{X} = (\mathbf{x}_1, \cdots \mathbf{x}_M)^T$ and $\mathbf{Y} = (\mathbf{y}_1, \cdots \mathbf{y}_N)^T$. Here $\mathbf{X}^T$ stands for the transposition of $\mathbf{X}$. Hereafter, we assume that nodes in subset V_x are located on the inner circle with radius $r_x = 1$, while nodes in V_y are located on the outer circle with radius $r_y = 2$. Note that $\|\mathbf{x}_m\| = 1$, $\|\mathbf{y}_n\| = 2$.

The centering (Young-Householder transformation) matrices are defined as $\mathbf{H}_M = \mathbf{I}_M - \frac{1}{M}\mathbf{1}_M\mathbf{1}_M^T$, $\mathbf{H}_N = \mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T$ where $\mathbf{I}_M$ and $\mathbf{I}_N$ stands for $M \times M$ and $N \times N$ identity matrices, respectively, and $\mathbf{1}_M$ and $\mathbf{1}_N$ are M- and N-dimensional vectors whose elements are all one. By using the double-centered matrix $\mathbf{B} = \{b_{m,n}\}$ that is calculated from the adjacency matrix $\mathbf{A}$ as $\mathbf{B} = \mathbf{H}_M\mathbf{A}\mathbf{H}_N$, we can consider the following objective function with respect to the position vectors $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_M)^T$ and $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_N)^T$.

$$S(\mathbf{X}, \mathbf{Y}) = \sum_{m=1}^{M}\sum_{n=1}^{N} b_{m,n} \frac{\mathbf{x}_m^T}{r_x}\frac{\mathbf{y}_n}{r_y} + \frac{1}{2}\sum_{m=1}^{M}\lambda_m(r_x^2 - \mathbf{x}_m^T\mathbf{x}_m) + \frac{1}{2}\sum_{n=1}^{N}\mu_n(r_y^2 - \mathbf{y}_n^T\mathbf{y}_n), \tag{5}$$

where $\{\lambda_m \mid m = 1, \cdots, M\}$ and $\{\mu_n \mid n = 1, \cdots, N\}$ correspond to Lagrange multipliers for the spherical constraints, i.e., $\mathbf{x}_m^T\mathbf{x}_m = r_A^2$ and $\mathbf{y}_n^T\mathbf{y}_n = r_B^2$ for $1 \le m \le M$ and $1 \le n \le N$. By maximizing $S(\mathbf{X}, \mathbf{Y})$ defined in Equation (5), we can obtain our visualization results, $\mathbf{X}$ and $\mathbf{Y}$ for a given bipartite graph.

2.3 Methods for Multigraph

For multigraphs, we denote the number of links from node u to v, i.e., (u, v), as $m_{u,v}$. Note that favoring or mentioning between a pair of users may occur several times during the observed period. We also denote an in-neighbor node set of node v by $A(v) = \{u; m_{u,v} \neq 0\}$, and an out-neighbor node set of node v by $B(v) = \{w; m_{v,w} \neq 0\}$. Then we can consider a node set $C(k) = \{v; |A(v)| = k\}$ for which the number of in-neighbor nodes is k, and a node set $D(k) = \{v; |B(v)| = k\}$ for which the number of out-neighbor nodes is k. Thus, by using these notations, with respect to the number of neighbors k, we can define the in-neighbor distribution $id(k)$ and the out-neighbor distribution $od(k)$ as follows:

$$in(k) = |C(k)|, \quad on(k) = |D(k)|. \tag{6}$$

Note that in case of simple directed graphs, the in- and out-neighbor distributions are simply called the in- and out-degree distributions, respectively.

Now, we define a set of nodes whose in-degree are not zero by $V^- = \{v \in V; deg^-(v) > 0\}$, and a set of nodes whose out-degree are not zero by $V^+ = \{v \in V; deg^+(v) > 0\}$.

Then, we can define the average in-multiplicity $m^-(v)$ for $v \in V^-$ and the average out-multiplicity $m^+(v)$ for $v \in V^+$ as follow:

$$m^-(v) = \frac{1}{|A(v)|}\sum_{u \in A(v)} m_{u,v}, \quad m^+(v) = \frac{1}{|B(v)|}\sum_{w \in B(v)} m_{v,w}. \tag{7}$$

For a multigraph, we can define the average in-multiplicity m^- and the average out-multiplicity m^+ as follow:

$$m^- = \frac{1}{|V^-|}\sum_{v \in V^-} m^-(v), \quad m^+ = \frac{1}{|V^+|}\sum_{v \in V^+} m^+(v). \tag{8}$$

On the other hand, with respect to number of neighbors $k(>1)$, we can define the average link multiplicity $im(k)$ for a node set $C(k)$, and the average link multiplicity $om(k)$ for a node set $D(k)$ as follows:

$$im(k) = \frac{1}{|C(k)|} \sum_{v \in C(k)} m^-(v), \quad om(k) = \frac{1}{|D(k)|} \sum_{v \in D(k)} m^+(v). \tag{9}$$

Similarly, for each node $v \in V$, we can define the in-Gini coefficient $g^-(v)$ for $v \in V^-$ and the out-Gini coefficient $g^+(v)$ for $v \in V^+$ as follow:

$$g^-(v) = \frac{\sum_{(u,x) \in A(v) \times A(v)} |m_{u,v} - m_{x,v}|}{2(|A(v)|-1) \sum_{u \in A(v)} m_{u,v}}, g^+(v) = \frac{\sum_{(w,x) \in B(v) \times B(v)} |m_{v,w} - m_{v,x}|}{2(|B(v)|-1) \sum_{w \in B(v)} m_{v,w}}. \tag{10}$$

For a multigraph, we can define the average in-multiplicity m^- and the average out-multiplicity m^+ as follow:

$$g^- = \frac{1}{|V^-|} \sum_{v \in V^-} g^-(v), \quad g^+ = \frac{1}{|V^+|} \sum_{v \in V^+} g^+(v). \tag{11}$$

With respect to number of neighbors $k(>1)$, we can define the average Gini coefficient $ig(k)$ for a node set $C(k)$, and the average Gini coefficient $og(k)$ for a node set $D(k)$ as follows:

$$ig(k) = \frac{1}{|C(k)|} \sum_{v \in C(k)} g^-(v), \quad og(k) = \frac{1}{|D(k)|} \sum_{v \in D(k)} g^+(v). \tag{12}$$

Here note that the gini coefficient has been widely used for evaluating inequality in a market [7]. We use this index to evaluate inequality between favoring and mentioning.

3 Summary of Data

We briefly explain the data we used in our analysis. These data are retrieved from Favorite, Follow, and Mention of Twitter.

"Favorites" is a function which enables users to bookmark tweets, or to browse them anytime. We constructed a network with the users as nodes, and the Favorer/Favoree relations as links. These data are retrieved from Favotter's "Today's best." [2] during the period from May 1st 2011 to February 12th 2012. Because of Favotter's specification, the retrieved tweets are bookmarked by more than or equal to 5 users. This directed network has 189,717 nodes, 7,077,070 simple links, and 33,456,690 multiple links [3].

[2] http://favotter.net/

[3] The number of simple links means that we count the multiple links between a pair of nodes as a single link.

"Follow" is the most basic function of Twitter. Users can get the new tweets posted by persons they are interested in by specifying whom to follow. We constructed a network with users who posted more than or equal to 200 tweets as nodes, and the follower/followee [3] relations as links. These data are retrieved from Twitter search [4] as of January 31st 2011. This directed network has 1,088,040 nodes and 157,371,628 simple links. Follow network does not have multiple links because users specify their respective followers only once.

"Mentions" are tweets which has the user's names of the form "@Screen_name" in the text. We constructed a network with users as nodes, and send/receive relations as links. These data are retrieved from Toriumi's data [8] for the period from March 7th 2011 to March 23rd 2011. This directed network has 4,565,085 nodes, 58,514,337 simple links and 193,913,339 multiple links.

Statistics of these networks are described for Tables 1 and 2. Here, WCC1 in Table 1 means the maximal weakly connected components, Em in table 2 means the number of multiple links. Others are defined in section 2.

Table 1 shows that Mentions network has a smaller WCC1 fraction than the other two networks. This is understandable in view of the communication aspect of Mentions because users do not send @-messages to people whom they do not well. Table 2 shows that Favorites network has smaller m^-, m^+, g^-, and g^+ (see equations 8 and 11) than Mentions. This is understandable because only a few users are heavy favorers and the majorities have much less favorees whereas in Mentions the distribution of the number of mentions of each user is less distorted, which makes the average degree of Mentions network larger than that of Favorites network.

Table 1: Statistics of simple directed networks

	$\|V\|$	$\|E\|$	$\|V\|_{WCC1}$ ($\|V\|_{WCC1}/\|V\|$)	d	c
Favorites	189,717	7,077,070	189,626 (99.9%)	37.3	0.2109
Follow	1,088,040	157,371,628	1,079,986 (99.3%)	144.6	0.7354
Mentions	4,565,085	58,514,337	1,839,189 (40.3%)	3.2	0.0387

Table 2: statistics of multi directed networks

	$\|V\|$	$\|Em\|$	d	m^-	m^+	g^-	g^+
Favorites	189,717	33,456,690	176.3505	2.1211	1.5024	0.2054	0.0851
Mentions	4,565,085	193,913,339	38.2894	3.6977	3.6574	0.3985	0.2138

[4] http://yats-data.com/yats/

4 Results

In this section, we report the results of analysis using various measures explained in 2.

4.1 Simple Directed Graph

As seen from Table 1, Favorites and Follow networks have each a large weakly connected component which includes almost all nodes but Mentions network is not so. Since Mentions network is too large to analyze for all nodes, we use WCC1 in the following analysis for Mentions network.

Degree Distribution Figures 2, 3, 4, 5, 6, and 7 are the results of degree distribution of the three networks. Blue and red diamond marks indicate *id* and *od* (see equation (3)), respectively. The vertical axis indicates the number of nodes in logarithmic scale. From these pictures, we see that all the networks can be said to have a scale-free property for both in-degree or out-degree.

Network Motif Figures 8 and 9 are the results of network motif analysis. The horizontal axis indicates the motif number explained in 4. In Figure 8 the vertical axis indicates the frequency of appearance in logarithmic scale, and in Figure 9 the vertical axis indicates z-score (see equation (4)) in logarithmic scale. Red and cyan bars mean positive score and negative score respectively. From these figures, we see that there are three predominant motifs: patterns 13, 12, and 8, which are all characterized by having mutual links, The results of Follow and Mentions networks are similar to these figures, so we omit showing these results.

4.2 Visualization of Bipartite Graph

Figure 10 is the result of visualization of bipartite graph of Favorites. In this analysis we used the data retrieved from only July 1st to 7th 2011 because so many links obscure the graph. Nodes on the outer circle are Favorers, and nodes on the inner circle are Favorees. Blue and Red nodes are users who are ranked Favorer/Favoree's top 10. Only links with more than or equal to 10 multiplicity are shown by gray lines.

NHK_PR is the official account of NHK's PR section[5], and sasakitoshinao is the account of freelance journalist. His tweets are on serious and important topics, for instance, current news or opinions about it. On the other hand, kaiten_keiku and Satomii_Opera are regular users of Twitter, and their tweets are often negative and/or "geeky".

From this figure, we see there are two clusters of popular users which are characterized by their content of tweets, one with serious and important tweets and the other with negative and/or geeky tweets.

[5] Japan Broadcasting Corporation

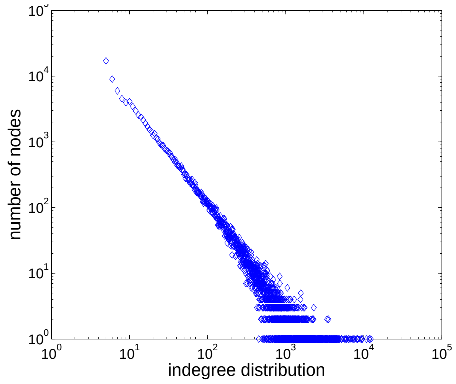

Fig. 2: Favorites network in-degree

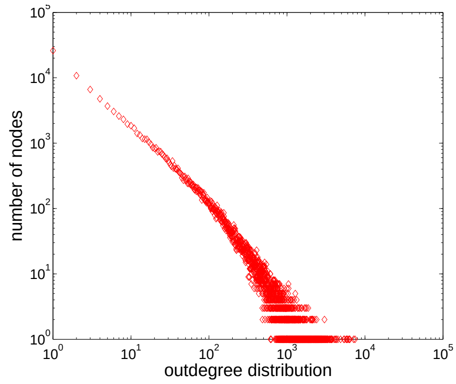

Fig. 3: Favorites network out-degree

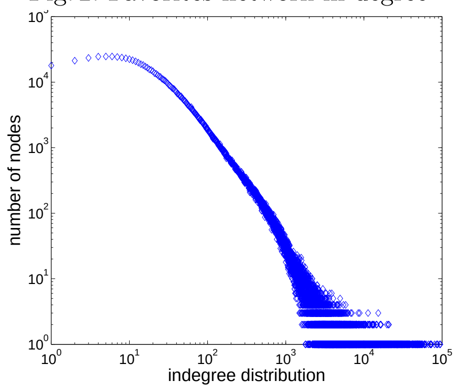

Fig. 4: Follow network in-degree

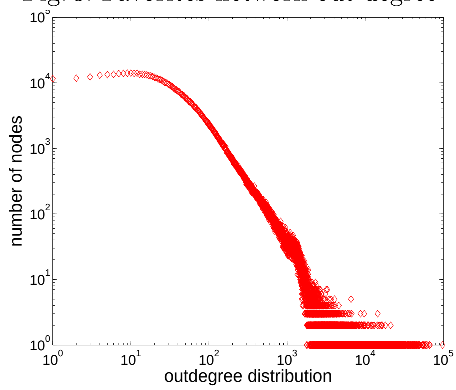

Fig. 5: Follow network out-degree

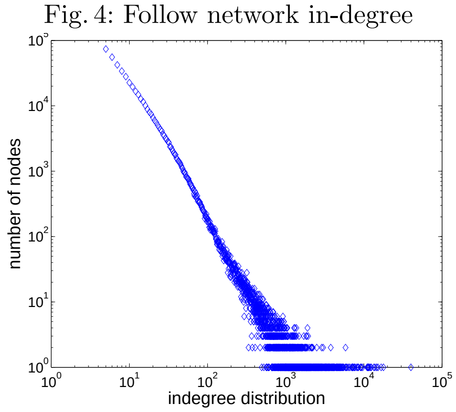

Fig. 6: Mentions network in-degree

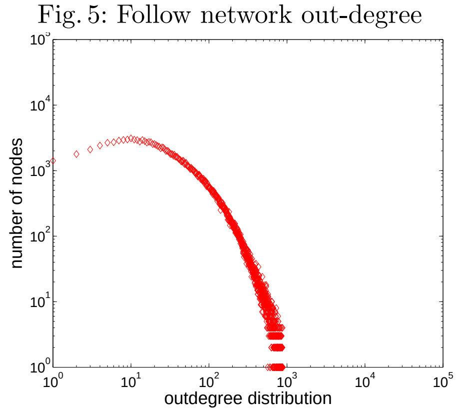

Fig. 7: Mentions network out-degree

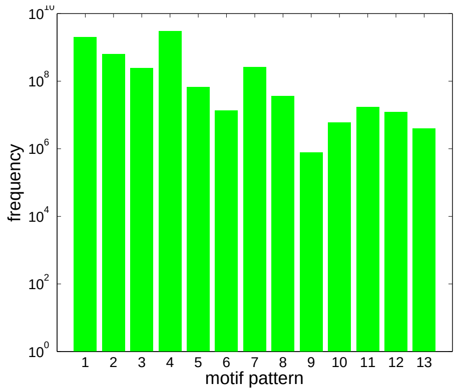

Fig. 8: Favorites network motif (frequency)

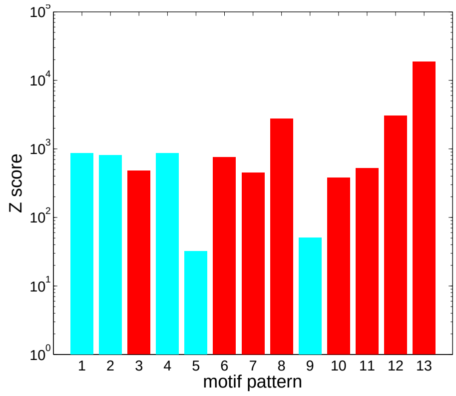

Fig. 9: Favorites network motif (z-score)

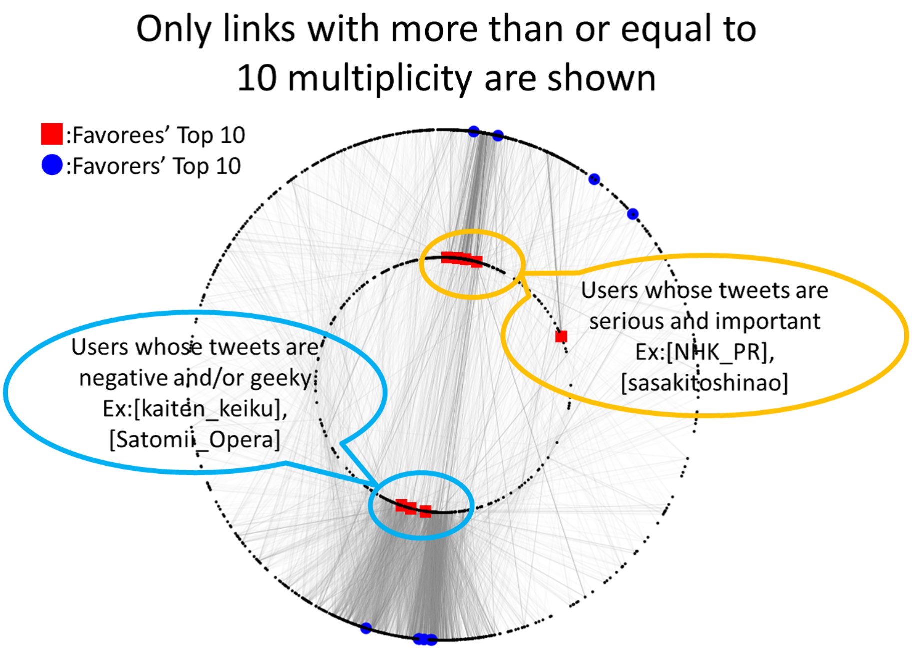

Fig. 10: Bipartite Graph Visualization

4.3 Multiple Directed Graph

In this subsection, we show the results of analysis using the measures explained in 2.3. In all the figures below (Figures 11 to 22), plots in blue squares are for in-degree, plots in red squares are for out-degree and plots in green circles are for randomized networks. Horizontal axes are all in logarithmic scale.

Degree Distribution Figures 11, 12, 13 and 14 are the results of degree distribution (see equation (6)) for Favorites and Mentions networks. The vertical axes are frequency (the number of nodes) in logarithmic scale. From these figures, we see that both networks have a scale-free property, same as the simple directed networks 4.1. We notice that the distributions for the randomized Mentions network are shifted right to the real Mentions network, but this is not so for Favorites network.

Average Multiplicity Figures 15, 16, 17 and 18 are the average multiplicity (see equation (7)) for the both networks. The vertical axes are in logarithmic scale. We notice the difference in correlation between the two networks. On the average, there are positive correlations between the average multiplicity and the degree for Favorites network (Figures 15 and 16), but the correlations change from positive to negative as the degree increases for Mentions network (Figures 17 and 18). Furthermore, the average multiplicity of randomized Favorites network behaves similarly to the real Favorites network, but that of randomized Mentions network is almost flat across all the range of degree.

Gini coefficient Figures 19, 20, 21 and 22 are the results of Gini coefficient (see equation (10) for the both networks. The vertical axes are in linear scale. Correlations between the Gini coefficient and the degree and the relation between the real and the randomized networks are similar to those for the average multiplicity, i.e., positive correlations for Favorites network (Figures 19 and 20), positive to negative correlations for Mentions network (Figures 21 and 22) and more positive correlations for the randomized Favorites network than the randomized Mentions network.

5 Discussion

The results in subsections 4.1 and 4.3 revealed that all the three networks have the scale-free property, but we notice that the variance in the degree distributions for Mentions network is smaller in high out-degree nodes than others. We conjecture that this is due to the communication aspect of Mention function, i.e. users do not send many @-messages to people they do not know well and, thus, there are probably no big hub nodes in Mentions network. Further, this also explains that the fraction of the maximal weakly connected component (defined in subsection 3) is smaller than the other networks.

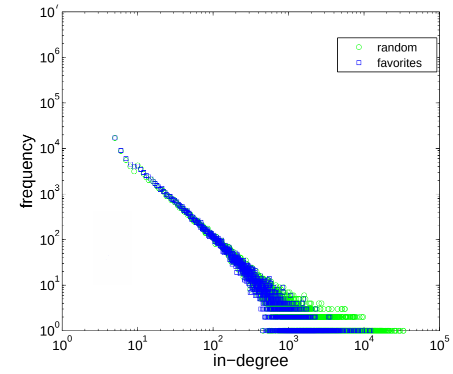

Fig. 11: Favorites in-degree

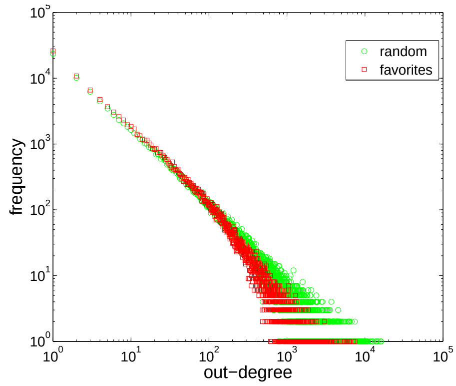

Fig. 12: Favorites out-degree

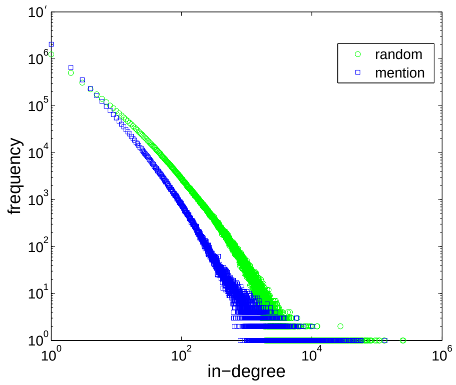

Fig. 13: Mentions in-degree

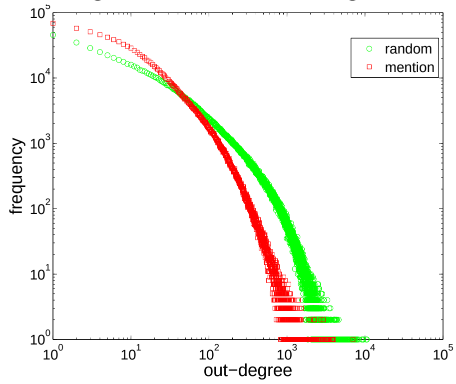

Fig. 14: Mentions out-degree

The results in subsection 4.1 revealed that there are a few numbers of predominant motifs that are characteristic of having mutual links. This accounts for the fact that, taking Favorites as example, mutual links are easily created between users who have similar tastes because Favorites network is driven by preference.

The results in subsection 4.2 that there are two clusters of popular users each corresponding to a particular type of tweets are quite natural and understandable. Whether these two are the unique tweets and there are no other such tweets remains to be explored.

The results in subsection 4.3 indicate that there are substantial difference in the distributions of multiplicity and Gini coefficient for high degree nodes between Favorites and Mentions networks. This is explainable considering the difference in nature of the two functions, Mentions network is driven by communications between users. Sending/receiving of @-message to/from many people become less practical, thus less frequent for high degree nodes. Favorites network is driven by preference. Expressing preference (bookmarking Favorees' tweets)

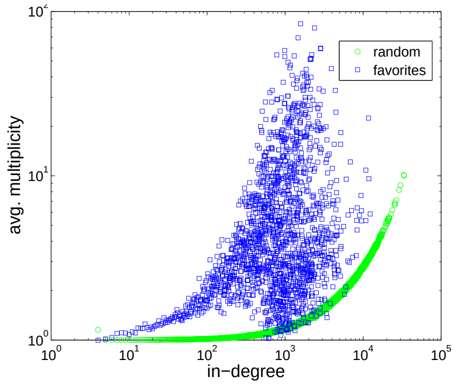

Fig. 15: Favorites in-degree

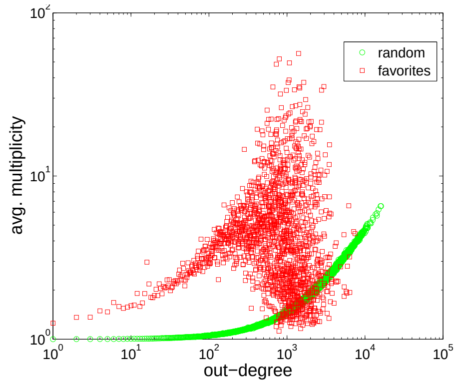

Fig. 16: Favorites out-degree

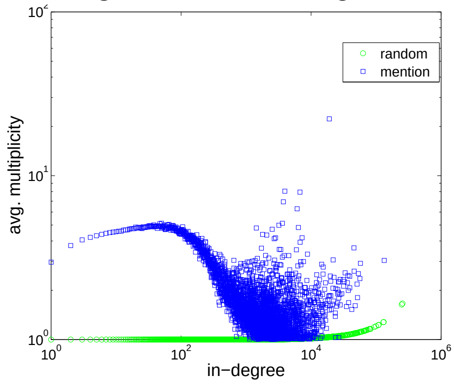

Fig. 17: Mentions in-degree

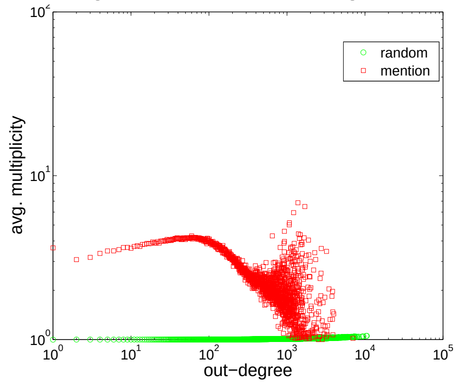

Fig. 18: Mentions out-degree

is much easier than sending/receiving message, thus relatively more frequent for high degree nodes.

The results in subsection 4.3 revealed that there are positive correlations between the Gini coefficient and the degree for all the range of degree for Favorites network, but not so for Mentions network. This may suggest that Favorers in high out-degree tends to preferentially bookmark specific Favorees' tweets, and vice versa for Favorees in high in-degree.

6 Conclusion

With the final goal of constructing a new user preference model in daily activities in mind, we analyzed, from the network structure perspective, the similarity and difference in the user behavior of the three functions of Twitter: Favorite, Follow and Mention. User behavior is embedded in the logs that users carried out these functions, which are represented by directed graphs. Favorite function was analyzed using three different graph representations: a simple graph, a

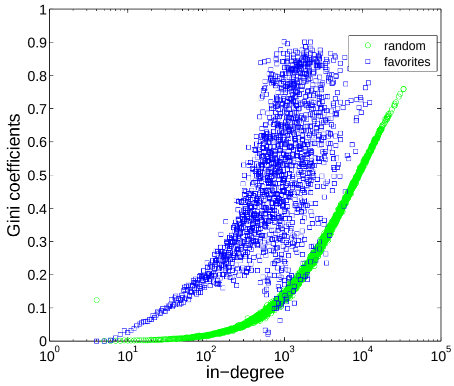

Fig. 19: Favorites in-degree

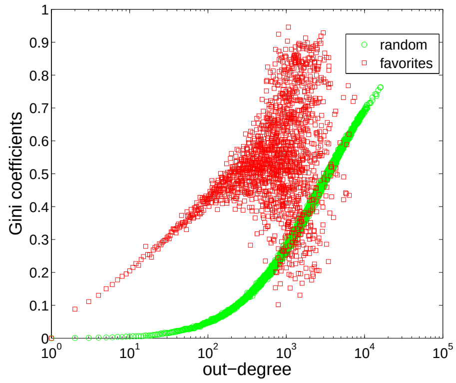

Fig. 20: Favorites out-degree

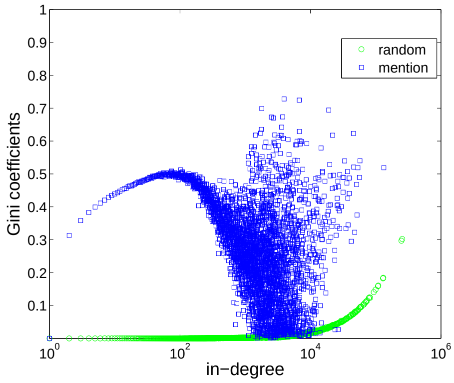

Fig. 21: Mentions in-degree

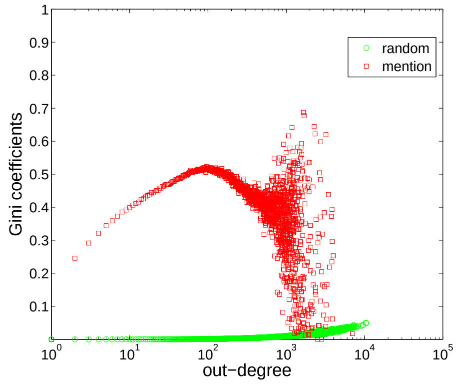

Fig. 22: Mentions out-degree

multigraph and a bipartite graph, Follow function by one graph representation: a simple graph, and Mention function by two graph representations: a simple graph and a multigraph. We used three real world Twitter logs to create these directed graphs and performed various kinds of analysis using several representative measures for characterizing structural properties of graphs, and obtained several salient features.

Major findings are that 1) Favorites and Followers networks have a very large connected component but Mentions network is not, 2) all the three networks (both simple and multiple) have the scale-free property in degree distribution, 3) all three networks (simple) have predominant three-node motifs having mutual links, 4) Favorites networks have freaks in Gini coefficient (one of the measures), 5) Favorites networks have two clusters of popular users, and 6) Favorites and Mentions networks differ in structure for high degree nodes in case of multigraph representation reflecting that Favorite operation is much easier than Mention operation although they are similar in case of simple graph representation.

As an immediate future work, we plan to obtain betweenness centrality, closeness centrality, or k-core percolation of Favorites network represented as a multigraph to further characterize use behavior and hopefully to extract enough regularity to model user preference, and pursue the literature review and usefulness of the model.

Acknowledgments

This work was partly supported by Asian Office of Aerospace Research and Development, Air Force Office of Scientific Research under Grant No. AOARD-11-4111, JSPS Grant-in-Aid for Scientific Research (C) (No. 22500133).

References

1. U. Brandes and T. Erlebach (Eds.), "Network analysis", LNCS 3418, pp. 293-317, Springer-Verlag, 2005.
2. T. Fushimi, Y. Kubota, K. Saito, M. Kimura, H. Motoda, and K. Ohara, "Speeding up bipartite graph visualization method", Proc. of the 24th Australasian Joint Conference on Artificial Intelligence (AI2011).
3. B.A. Huberman, D.M. Romero and F. Wu, "Social networks that matter: Twitter under the microscope", First Monday, Vol. 14, No. 1, 2009.
4. H.Kwak, C.Lee, H.Park, and S.Moon, "What is Twitter, a social network or a news media?", In Proceedings of the 19th international conference on World Wide Web, pp.591-600, 2010.
5. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii and U. Alon, "Network motifs: simple building blocks of complex networks", Science 298, pp. 824–827, 2002.
6. M. E. J. Newman,"The structure and function of complex networks", SIAM Review, Vol.45, pp.167–256, 2003.
7. M.J. Salganik, P.S. Dodds, and D.J. Watts. "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market", Science 311, pp. 854–856, 2006.
8. F. Toriumi, K. Shinoda, S. Kurihara, T. Sakaki, K. Kazama, I. Noda,"Disaster Changes Social Media", In Proceedings of the 7th Conference of JWEIN, 2011.