Real-World Popularity Estimation from Community Structure of Followers on SNS

Shuhei Kobayashi
Kyoto University
Kyoto, Japan
skobayashi@dl.soc.i.kyoto-u.ac.jp

Keishi Tajima
Kyoto University
Kyoto, Japan
tajima@i.kyoto-u.ac.jp

***Abstract*—In this paper, we propose methods of estimating the real-world popularity of users of online social network services (SNSs). Because their followers on an SNS are biased sampling from their real-world fans, we cannot estimate their real-world popularity simply by the number of their online followers. Our methods are based on the following hypothesis: SNS users with followers more distributed over the SNS graph are likely to have more real-world popularity. Because the entire social graph is often unavailable, we design four methods of measuring how much followers are distributed by using only the local structure of the neighbors of the followers. Three of them uses variations of the clustering coefficient of node, and one of them uses a metric we newly designed. Through the development and evaluation of our methods, we validate the hypothesis above.**

***Index Terms*—social network, Twitter, Instagram, clustering coefficient**

I. Introduction

On today's online social network services (SNSs), such as Twitter and Instagram, a wide range of people, from super famous celebrities to nameless ordinary people, can obtain popularity by posting messages on some topics. The number of followers is the most widely used metric for measuring the popularity of such SNS users. The number of followers generally works well as a metric of popularity, but there are several factors that can make it inaccurate. For example, the existence of bought fake followers is a well known problem. For this problem, there have been studies on the detection of such fake accounts without real persons behind them. The definition of the popularity also affect the accuracy because we may not define popularity simply by the number of fans.

However, even when we do not have fake users and we define the popularity by the number of fans, the number of followers is not still a perfect metric of popularity. It is because the set of followers of some user's SNS account is a sampled set of the fans of the user, and the sampling process can involve various kinds of bias. Sometimes most fans of an user explicitly follow the user's account, and sometimes there are many fans of a user who do not explicitly follow the user's account. This discrepancy becomes especially evident when we use the number of followers to estimate a user's popularity in the real-world, not the online popularity of the account.

One of the factors that make this discrepancy is the explicit request of the "follow" actions from the account user. That is, sometimes SNS users explicitly ask people within the close community, e.g., their personal friends, to follow their account. Sometimes people within the same community as an SNS user voluntarily do it even if the SNS user does not ask it. In such cases, these accounts will have a larger number of followers than users who have similar degree of popularity but do not explicitly request it.

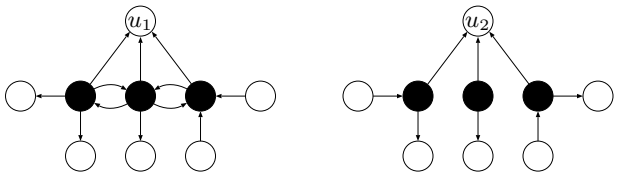

Fig. 1. Example of users with the same number of followers but with different community structure of the followers. Both u_1 and u_2 have three followers. Followers of u_1 are densely connected, which means u_1 is followed only by people in one community, while followers of u_2 are unrelated with each other, which means u_2 may have popularity in wider range of people.

The example above is an extreme case, but in general, if two SNS accounts u_1 and u_2 have the same number of followers, and the followers of u_1 are densely connected with each other while the followers of u_2 are distributed over the social graph without dense connection, we expect that u_1 has many followers densely sampled from the close community of u_1, and the user u_2 actually has more fans. Because the followers of u_2 are distributed over the graph, u_2 has a greater chance of having many latent fans within the neighbors of those followers. Figure 1 illustrates this example. Both u_1 and u_2 have three followers. Followers of u_1 are densely connected, which means u_1 is followed only by people in a local community, while followers of u_2 are unrelated with each other, which means u_2 has more global popularity.

Following this observation, in this paper, we propose and compare several methods of estimating the real-world popularity of SNS users based not only on the number of their followers but also on their community structure, i.e., the structure formed by the follow edges from/to the followers.

Because our methods use that information, our methods need to collect the following data for estimating the popularity of a given target user:

- the list of followers of the given target user,
- the list of the followers of each follower, and
- the list of the followees (friends) of each follower.

By collecting these three types of data, we construct a subgraph of the SNS's social graph induced by these users. We call this graph *the follower-neighbor graph* of the target user.

One advantage of this approach is that we do not need the information on the entire social graph of the SNS. There have been many studies proposing methods for estimating some properties of nodes in a large graph, and they can be classified into two types: those that require the entire graph and those that require only the local structure in the neighbor of the node in question. Many (not all) SNSs allow us to retrieve local graph structure of some nodes, but today's most SNSs restrict large scale crawling of their social graphs. For example, Twitter is selling their tweet information, but do not sell their social graph data. Instagram allows us to browse the list of followers of a given user, but they forbid scraping of the follower lists. As a result, the entire social graph is not available for most of us. Therefore, methods that only require the local structure of a node in question are advantageous.

For each follower, the following types of information can be obtained only from the follower-neighbor graph without requiring the entire graph:

- how many followers each follower has,
- how many followees (friends) each follower has, and
- which pairs of the followers are connected and which pairs are not.

our methods first calculates the score of each follower of the target user by using these three types of information. We then sum up the scores of all the followers to estimate the popularity of the target user.

For calculating the score of each follower, we propose four methods. Three of them uses the clustering coefficient and its variations. The clustering coefficient is a widely used metric for measuring the edge density within the neighbor of a node in a graph. If the clustering coefficient of a node is large, the node is mainly connected to a small number of specific communities, and if the clustering coefficient of a node is small, the node is connected to wider communities distributed over the graph. We expect that a user with followers with smaller clustering coefficients have greater real-world popularity. Therefore, we give higher scores to the followers with smaller clustering coefficients.

The fourth methods define the score of the followers more directly following our hypothesis. For each follower, we hypothesize that the larger the number of followers (i.e., 2-hop followers of the target user) is, the greater the follower's influence is, and the larger the number of followees (friends) is, the wider the range of the communities the follower is connected to, and also the smaller the likelihood that the follower is a personal friend of the target user. In summary, the larger the number of followers and followees of a follower is, the greater the value of the follower.

We conducted experiments for evaluating and comparing our proposed methods and a simple baseline method that only uses the number of followers as the metric of the popularity. For the evaluation, we need a dataset with the ground truth of the real-world popularity. To collect such dataset, we focused on Ms/Mr university competitions. We chose 9 universities that have competitions where the winner and the runners-up are chosen not solely by the judges of the competition but based on the public popularity votes (either online, offline on the contest day, or both). All the contestants of those competitions have their Twitter accounts. We collected follower-neighbor graphs of the contestants' Twitter accounts, estimated their popularity by our methods and by the baseline method, and evaluate their accuracy based on the results of the competitions.

Through the development and evaluation of these methods, we validate the following two hypothesis:

- SNS users with followers more distributed over the SNS graph are likely to have more real-world popularity, and therefore, we can improve the accuracy of the popularity estimation by using the information on the community structure of the followers compared with the simple method solely based on the number of followers.

The remainder of this paper is organized as follows: Section II describes related research, Section III details the proposed method, Section IV describes experimental details, and Section V concludes the paper.

II. Related Work

In this section, we first survey the research on the estimation and prediction of the popularity of SNS users. We also explain existing metrics on the edge density in the neighbors of a node in a graph. The most well-known method of measuring it is the clustering coefficient [1]. We briefly survey research on the extensions of the clustering coefficient. We also explain a study that use the clustering coefficient in a problem related to our problem.

A. SNS user popularity prediction

Although there have been a large number of studies on the prediction of the popularity of online contents, such as tweets, images on Flickr, and videos on YouTube, there have been only a few research on the prediction or estimation of the popularity of SNS users because most studies define the popularity of users by the number of followers or by the popularity of their posts in the past, not by their real-world popularity.

Imamori and Tajima [2] proposed a method of predicting the popularity that a new Twitter user will have in future. Osawa and Matsuo [3] proposed a method of predicting the popularity that some real-world entity, e.g., a person, will have if it creates an SNS account. On the other hand, our purpose is the opposite of the latter one: estimation of the real-world popularity of an entity from the information of its SNS account.

There have been more research on the estimation of the influential power of SNS users [4]–[7]. However, Romero et al. [5] has shown that popularity and influence power do not coincide. In addition, these studies estimate the SNS users' influential power on SNS, not the estimation of offline real-world popularity from the online information.

B. Clustering coefficients for weighted directed graphs

Suppose we have an undirected unweighted graph $G(V, E)$ where V is the set of nodes and E is the set of edges. The clustering coefficient of a node $v_i \in V$, denoted by $C(v_i)$, indicates the degree to which its neighbors are connected to each other, and is defined as below:

$$C(v_i) = \frac{2|\{\{v_j, v_k\} \in E \mid \{v_i, v_j\} \in E, \{v_i, v_k\} \in E\}|}{k_i(k_i - 1)}$$

where k_i is the degree of v_i. That is, $C(v_i)$ finds the pair of v_i's neighbors forming a triangle with v_i, and calculate the ratio of the number of existing triangles to the number of all possible pairs.

Various extensions of clustering coefficients have been proposed, some of which are summarized by Saramäki et al. in [8]. In the following, we explain some of the extensions including those we use in our methods.

There are many applications where we need to model networks as weighted graphs. In weighted graphs, a weight representing the strength of the connection is assigned to each edge. When we analyze such data, we sometimes want to take into account the weights in the calculation of clustering coefficient. Barrat et al. [9] proposed clustering coefficients for weighted graphs. They define the weight of a triangle by the sum of the weights of the two edges adjacent to the node in question. On the other hand, the method proposed by Onnela et al. [10] defined clustering coefficients for weighted graphs by the formula below:

$$\tilde{C}(v_i) = \frac{1}{k_i(k_i - 1)} \sum_{j,k} (\hat{w}_{ij}\hat{w}_{ik}\hat{w}_{jk})^{1/3}$$

In the formula, $\hat{w}_{ij}$ is the normalized weight defined by $\hat{w}_{ij} = w_{ij} / \max_{x,y}(w_{xy})$ where w_{ij} is the weight of the edge connecting v_i and v_j. Because they use the geometric mean of the weights of edges in a triangle, if there is an edge with a small weight in a triangle, its contribution to the clustering coefficient is very small. The $\tilde{C}(v_i)$ is 1 if all adjacent nodes of v_i are connected and the weights of the edges in the triangles equal to $\max_{x,y}(w_{xy})$, the maximum weight in the graph.

Fagiolo [11] has proposed clustering coefficients for directed graphs. For unweighted directed graphs, their clustering coefficients are defined by the ratio of existing directed triangles to the total number of all possible directed triangles. In other words, they distinguish triangles consisting of the same nodes but consisting of edges in different directions. For weighted directed graphs, they extended the clustering coefficient for weighted graphs by Onnela et al. [10] in the same manner by distinguishing the triangles with different edge directions. In this study, we use their definition of clustering coefficients for weighted directed graphs.

C. Use of clustering coefficient in a related problem

Berahmand et al. [12] used clustering coefficient and the second-level clustering coefficient for estimating the influential power of a node in a graph by using only semi-local information, i.e., without global graph data. However, their aim is to estimate influential power, not popularity, and the usage of the clustering coefficient is completely different from ours.

III. Proposed Method

In this section, we describe our methods of estimating the real-world popularity of SNS users. As explained in Section I, our method estimate it based on whether the user in question has the followers from wider communities and is expected to have global popularity, or it has followers only from a small local community, and seems to only have local popularity.

Our methods first calculates a score of each follower of the target user, which represents how much the follower is globally connected to wider communities. We then sum up the scores of all the followers, and use the result as the metric for estimating the real-world popularity of the target user.

We first explain three methods that assign scores to the followers based on their clustering coefficients. When many followers are connected to each other, it is highly likely that they form a group, and the target user has followers mainly from a small local communities. On the contrary, if the followers are not densely connected with each other, they are less likely to form a group and expected to be connected to wider communities. Therefore, if the clustering coefficient of the followers are large, we expect that the real-world popularity of the user is relatively low compared with other users with the similar number of followers. If the clustering coefficient of the followers are small, we expect the opposite.

Because a clustering coefficient takes its maximum value 1 when the node is embedded in a dense local community, we subtract the clustering coefficient from 1, and the resulting value is used as the score of the node.

A. Scores based on undirected clustering coefficient

Our simplest method based on the clustering coefficient uses the most classic clustering coefficient for unweighted undirected graphs explained in Section II. It is well known that the clustering coefficients of nodes in social graphs are high [1]. It is one of the most important characteristic of social networks, and has also been observed in SNSs [13].

Even though the clustering coefficients of nodes in social graphs are generally high, there is still a large variation among the nodes. We use it to estimate the expected contribution of each follower to the popularity of the target user.

B. Scores based on directed clustering coefficient

We also compare a method that calculates the scores of nodes based on the clustering coefficient for directed graphs. We follow the definition by Fagiolo [11].

As explained in Section II, it is defined by the ratio of the existing directed triangles to the all possible directed triangles. By using it, we can take into account more information available in the social graph, which is a directed graph. The value of a directed clustering coefficient becomes larger when related nodes are connected reciprocally than when they are connected only by one-way edges. Therefore, we can give higher coefficients (and therefore, lower scores) when related nodes are more strongly connected.

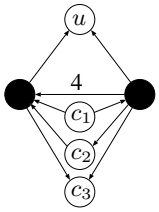

Fig. 2. Weights of edges between followers of the target user u. Two filled circles are followers of u. u is always their common neighbor, and in this example, they have three other common neighbors, c_1, c_2, c_3. Therefore, they have four common neighbors in total, and the weight of the edge between them is defined to be 4. Note that the weight is always larger than 1 because the target user is always a common neighbor for its followers.

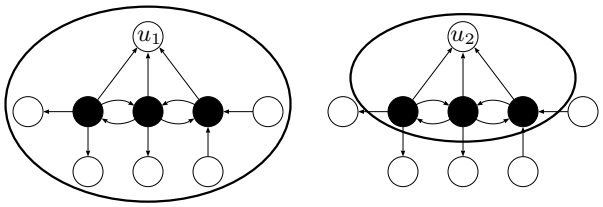

Fig. 3. Subgraphs used for the calculation of clustering coefficients. Two graphs are the follower-neighbor graphs of target users u_1 and u_2. Filled circles represent the direct followers of the target users, and nodes at the bottom of each graph are followers/followees of the direct followers. The graph on the left depicts the subgraph used in the calculation in the first and second methods that use the classic clustering coefficient and the directed clustering coefficient. In those methods, we use the entire follower-neighbor graphs. The graph on the right depicts the subgraph used in the calculation in the third method that uses the weighted directed clustering coefficient. In that method, we use subgraphs within the circle which only includes the target user and its direct followers.

C. Scores based on weighted directed clustering coefficient

The discussion above suggests that the strength of the connection among related nodes are useful information. Therefore, we next consider a method that use the clustering coefficient for weighted directed clustering.

There are several ways to define the strength of the connection between two nodes by only using the graph structure, but the most popular one is the common neighbor method. We adopt it in our method. The strength of the connection between two followers is considered stronger when they have more followers/followees in common.

Figure 2 illustrates the idea by using an example. In this example, two filled circles are the followers of the target user u. Because we only consider followers of the target user, they always have at least one common neighbor, which is the target user. In this example, the two followers also have three other common neighbors c_1, c_2, c_3. Therefore, they have four common neighbors in total, and we define the weight of the edge between the two followers to be 4.

On the other hand, if there is not an edge between two followers, the strength of their connection is 0 even if they have many common neighbors. In summary, the weight of the edge between two followers v_1 and v_2, denoted by w_{ij}, is defined as follows:

$$w_{ij} = \begin{cases} \text{the number of common neighbors} & \text{if edge exists} \\ 0 & \text{otherwise} \end{cases}$$

When we calculate clustering coefficients of followers, we also need the weights of edges connecting the followers and the target user. We set it to be the average of the weights of the edges between the followers.

We then calculate the clustering coefficients of the followers by using these edge weights. When calculating them, we only include the target user and its followers in the graph, and exclude all other nodes from the graph. Therefore, the meaning of the clustering coefficient in this method is different from the meaning of the clustering coefficient in the two previous methods. In this method, clustering coefficients of the followers are calculated only based on their connection with the target user and with each other.

The difference is depicted in Figure 3. Each of the two subgraphs represents our follower-neighbor graph. The nodes u_1 and u_2 at the top of each graph are target nodes, filled circles represent the direct followers, and nodes at the bottom of each graph are followers/followees of the direct followers. The graph on the left depicts the subgraph used in the calculation in the first and second methods that use the classic clustering coefficient and the directed clustering coefficient. In those methods, we use the entire follower-neighbor graphs. The graph on the right depicts the subgraph used in the calculation in the third method that uses the weighted directed clustering coefficient. In that method, we use subgraphs within the circle which only includes the target user and its direct followers.

The clustering coefficients of nodes in the produced weighted directed graphs are calculated following the definition by Onnela et al [10], explained in Section II.

D. Scores based on the number of edges between followers and the number of 2-hop followers/followees

The previous three methods use variations of the clustering coefficient. Our fourth method tries to define the metric directly following our hypothesis.

For a given follower, the more followers (i.e., 2-hop followers of the target user) it has, the more influential it is, and the more followees it has, the less likely it is a personal friends of the target user. In addition, the greater the number of followers and followees are, the wider the community the follower is connected to. Therefore, our hypotheses is that followers with the larger number of followers and followees are expected to more contribute to the popularity of the target user.

In addition, even if the followers of the target user have many followers and followees, if the followers are also closely connected to each other, the communities to which the followers are connected are likely to be limited, and it is a negative factor in the estimation of the popularity of the target user.

Figure4 illustrates the idea. The red nodes are the direct followers of the target node u, and the blue nodes are the followers/followees of the direct followers. We hypothesize that the larger the number of blue nodes is, the greater the

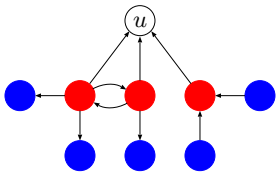

Fig. 4. The overview of our fourth method. The red nodes are the direct followers of the target node u, and the blue nodes are the followers/followees of the direct followers. We hypothesize that the larger the number of blue nodes is, the greater the popularity of u is, and the smaller the number of edges between red nodes is, the greater the popularity of u is. To represent the latter, for each follower (a red node), we count the number of connected red nodes.

popularity of u is, and the smaller the number of edges between red nodes is, the greater the popularity of u is.

Following these observations, we define the score of each follower in our fourth method by the formula below:

$$\text{score} = \frac{\log_2(\#\text{ of non-red nodes adjacent to red nodes} + 2)}{\log_2(\#\text{ of red nodes adjacent to red nodes} + 2)}$$

By taking the inverse of the logarithm of the number of red nodes adjacent to red nodes, the more adjacent red nodes the follower has, the smaller the score of the follower is. In addition, by the logarithm of the number of non-red nodes adjacent to red nodes (i.e., blue nodes), the more blue nodes the follower has, the larger the score of the follower is.

The reason for taking the logarithm is the number of adjacent nodes usually follow power law distribution. We add 2 to the number of adjacent nodes to prevent the scores from diverging or taking negative values when the number of adjacent nodes is 0 or 1.

E. Final Score of the target user

Given the score of each follower, we sum up their scores and the result is used as the score of the target user. Because we sum up the score of the followers, generally speaking, the more follower the target user has, the higher the score of the target user is. However, the followers are not count equally. They are given scores, which are used as their weight.

In the first three methods that use clustering coefficients, the score is defined to be $1-$ coefficient, as explained before. Therefore, the score takes the value between 0 and 1, and the score is bigger when the clustering coefficient is small.

A clustering coefficients cannot be defined for locked accounts because we cannot obtain the list of their followers and the list of their followees. When a direct follower of the target node is locked, we assign the average score value of the open (i.e., not locked) direct followers of the target node.

A clustering coefficient of a follower cannot be defined also when the follower only has one adjacent node. Note that they always have at least one adjacent node because they are followers of the target node. For such followers, we also use the average score of the other follower nodes. However, a follower with only one adjacent node is very rare.

Similarly, in the fourth method, if a follower is a locked account, we use the average number of adjacent red nodes and the average number of adjacent blue nodes of the other open followers.

IV. Experiments

In this section, we describe the details of the experiment we conducted, and discuss the result.

We conducted experiments by using the data on Ms/Mr competitions of the following nine universities in Japan: Komazawa University, Gakushuin University, Ryukoku University, Chuo University, Seikei University, University of Tokyo, Kwansei Gakuin University, Doshisha University, and Kansai University.

In the experiment, we collected information on the Twitter accounts of the female contestants of the competitions of those universities. Data were retrieved on the day of the competition or the day before that for most universities. For some universities, data were retrieved several days after the competition, but the number of followers did not change significantly during those several days. Therefore, we believe that this does not largely affect the result of the experiment.

We then construct the follower-neighbor graph for each contestant, estimate the popularity of each contestant by using our method, and predict the winner and the runners-up of each competition. We then compare the result of our prediction and the result of the competitions, which are determined based on the public popularity.

We also compare the accuracy of our methods with the accuracy of a simple baseline method. The baseline method predicts the winner and the runners-up simply based on the number of followers of the Twitter accounts of the contestants.

A. Evaluation metrics

We evaluate the accuracy of our methods and the baseline method by using the following metrics::

1) **combination**: accuracy where the prediction is regarded as correct if the winner and the runner-up of the competition are the top-two contestants in the prediction,
2) **ordered**: accuracy where the prediction is regarded as correct if the winner is the top contestant in the prediction, and the runner-up is the second best contestant in the prediction,
3) **winner only**: accuracy where the prediction is regarded as correct if the winner is the top contestant in the prediction, and
4) **separate**: accuracy where the prediction of the winners and the prediction of the runner-up are counted separately. For example, if a method correctly predicted the winners of 3 competitions and the runners-up of 4 competitions, the accuracy of the method is $(3+4)/(9+9) = 7/18$.

B. Results

Tables I through IX show the results of the prediction by the baseline method, which simply uses the number of followers on Twitter, and our four prediction method, denoted by s1 to s4, for nine competitions. Each score is rounded down to the

TABLE I
KOMAZAWA

	# followers	s1	s2	s3	s4	result
No1	1900	1687	1707	1892	3713	2nd
No3	3126	2826	2835	3115	5872	1st
No5	1371	1235	1241	1359	2563	
No7	1320	1129	1158	1309	2750	

TABLE II
GAKUSHUIN

	# followers	s1	s2	s3	s4	result
No1	2489	2149	2167	2475	4607	
No2	2577	2385	2389	2568	5223	2nd
No3	2836	2536	2548	2827	4929	
No4	2694	2374	2393	2685	5211	
No5	3212	2931	2943	3210	6373	1st

TABLE III
RYUKOKU

	# followers	s1	s2	s3	s4	result
No1	1680	1546	1549	1670	2732	
No2	2520	2341	2344	2515	4004	
No3	1114	1016	1018	1108	1642	
No4	1767	1582	1589	1754	3234	1st
No5	1165	1052	1058	1157	2110	2nd
No6	898	778	790	889	2048	

TABLE IV
CHUO

	# followers	s1	s2	s3	s4	result
No1	1785	1444	1508	1776	3729	
No2	1797	1393	1477	1788	3763	
No3	2813	2331	2403	2801	7244	1st
No4	1720	1433	1481	1700	4134	2nd
No5	1563	1306	1337	1551	2916	

TABLE V
SEIKEI

	# followers	s1	s2	s3	s4	result
No1	2921	2662	2672	2914	5422	
No2	2416	2159	2165	2398	4403	2nd
No3	3411	2969	2985	3395	5549	1st
No4	938	810	817	919	1837	
No5	1310	1067	1089	1300	2709	

TABLE VI
TOKYO

	# followers	s1	s2	s3	s4	result
No1	2941	2566	2598	2930	5863	
No2	4611	3536	3650	4597	8754	2nd
No3	2096	1658	1687	2085	3607	
No4	3201	2865	2922	3128	9398	1st
No5	2275	2050	2058	2266	4454	

nearest integer number. The column labeled "result" shows the true results of the competitions. Each row corresponds to one contestant, and it shows the number of followers and the score given by our four methods. Yellow cells are the top contestant in each prediction or in the true result, and blue cells are the second best contestant in each prediction or in the true result.

Table I shows that all the baseline and the proposed methods correctly predicted the winner and the runner-up for the competition of Komazawa University. For the competition of Gakushuin University (Table II), only our fourth method correctly predicted both the winner and runner-up, and all

TABLE VII
KWANSEI GAKUIN

	# followers	s1	s2	s3	s4	result
No1	3856	3430	3446	3845	6511	
No2	3721	3380	3388	3713	7060	
No3	869	716	746	849	2000	
No4	4010	3616	3632	4000	6748	1st
No5	1939	1712	1725	1926	3752	
No6	3660	3354	3363	3647	6315	2nd

TABLE VIII
DOSHISHA

	# followers	s1	s2	s3	s4	result
No1	4994	4014	4159	4979	12170	1st
No2	1310	1093	1118	1299	2596	
No3	3785	3004	3089	3769	7601	
No5	2400	2015	2047	2387	4618	2nd
No6	2471	2184	2202	2458	4521	

TABLE IX
KANSAI

	# followers	s1	s2	s3	s4	result
No1	2919	2487	2511	2907	5674	1st
No2	2366	2170	2176	2353	3985	
No3	3694	3420	3432	3689	7345	2nd
No4	2378	2147	2160	2372	4670	
No5	1865	1671	1683	1854	3513	
No6	1662	1517	1523	1651	2867	

TABLE X
SUMMARY

	combination	ordered	winner only	separate
# followers	3/9	1/9	6/9	7/18
s1	3/9	1/9	6/9	7/18
s2	3/9	1/9	6/9	7/18
s3	3/9	1/9	6/9	7/18
s4	5/9	4/9	6/9	10/18

the other methods correctly predicted the winner, but wrongly predicted the contestant No 3 as the runner-up while the true runner-up was the contestant No 2.

Table X summarizes the total accuracy of the baseline method and each of the proposed methods over the nine competitions. It shows that all the methods are tie for winner-only accuracy, but our fourth method was the best for all the other three accuracy metrics we defined before. This result validates our two hypotheses that was explained before and shown again below:

- SNS users with followers more distributed over the SNS graph are likely to have more real-world popularity, and therefore, we can improve the accuracy of the popularity estimation by using the information on the community structure of the followers compared with the simple method solely based on the number of followers.

On the other hand, the accuracy of the other three proposed methods that use variations of the clustering coefficient were exactly the same as the accuracy of the baseline method. In more detail, the prediction by our third method, which uses the weighted directed clustering coefficient, was exactly the same as the prediction by the baseline method, and the prediction by the other two methods, s2 and s3, are also the same as the prediction of the baseline method except for the prediction

TABLE XI
KOMAZAWA:TWITTER+INSTAGRAM

	Twi	Ins	sum	s1	s2	s3	s4	result
No1	1900	3385	5285	2446	2475	2743	5382	2nd
No3	3126	4882	8008	3941	3953	4343	8186	1st
No5	1371	2154	3525	1725	1734	1897	3579	
No7	1320	5172	6492	2246	2303	2604	5471	

TABLE XII
GAKUSHUIN:TWITTER+INSTAGRAM

	Twi	Ins	sum	s1	s2	s3	s4	result
No1	2489	3354	5843	2880	2903	3317	6174	
No2	2577	4920	7497	3534	3540	3806	7740	2nd
No3	2836	2283	5119	3051	3065	3402	5931	
No4	2694	3885	6579	3238	3265	3662	7108	
No5	3212	3912	7124	3832	3848	4197	8333	1st

TABLE XIII
RYUKOKU:TWITTER+INSTAGRAM

	Twi	Ins	sum	s1	s2	s3	s4	result
No1	1680	1612	3292	1921	1924	2074	3394	
No2	2520	2585	5105	2947	2951	3166	5041	
No3	1114	1740	2854	1417	1419	1545	2290	
No4	1767	3614	5381	2399	2410	2659	4904	1st
No5	1165	1781	2946	1458	1466	1604	2924	2nd
No6	898	1881	2778	1190	1208	1360	3130	

TABLE XIV
CHUO:TWITTER+INSTAGRAM

	Twi	Ins	sum	s1	s2	s3	s4	result
No1	1785	6677	8462	2808	2932	3452	7249	
No2	1797	2500	4297	1882	1996	2415	5085	
No3	2813	14625	17438	5390	5558	6477	16749	1st
No4	1720	1832	3552	1818	1879	2157	5245	2nd
No5	1563	6092	7655	2592	2653	3077	5784	

TABLE XV
SEIKEI:TWITTER+INSTAGRAM

	Twi	Ins	sum	s1	s2	s3	s4	result
No1	2921	2208	5129	3171	3182	3470	6456	
No2	2416	3144	5560	2868	2876	3186	5849	2nd
No3	3411	5172	8583	4105	4128	4694	7672	1st
No4	938	948	1886	1017	1025	1153	2306	
No5	1310	1816	3126	1440	1470	1755	3657	

TABLE XVI
TOKYO:TWITTER+INSTAGRAM

	Twi	Ins	sum	s1	s2	s3	s4	result
No1	2941	1440	4381	2883	2919	3292	6588	
No2	4611	6315	10926	4758	4912	6186	11779	2nd
No3	2096	1989	4085	2055	2091	2585	4471	
No4	3201	3130	6331	3573	3644	3900	11717	1st
No5	2275	852	3127	2244	2252	2481	4875	

for the competition of Chuo University. These results show that variations of the clustering coefficient is not large enough to produce the prediction which is different from the simple prediction by the number of followers.

C. Additional Use of Instagram Information

All the contestants involved in the previous experiment also have Instagram account. Therefore, we could apply our methods also to their Instagram accounts for predicting the competition results. However, Instagram does not provide API for accessing the social graph information, and also forbid the scraping of the information. Therefore, we could not retrieve

TABLE XVII
KWANSEI:TWITTER+INSTAGRAM

	Twi	Ins	sum	s1	s2	s3	s4	result
No1	3856	6937	10793	4988	5010	5591	9467	
No2	3721	5524	9245	4647	4657	5104	9705	
No3	869	1759	2628	1082	1127	1283	3022	
No4	4010	9503	13513	5779	5805	6392	10785	1st
No5	1939	2471	4410	2263	2280	2545	4959	
No6	3660	9336	12996	5513	5529	5995	10381	2nd

TABLE XVIII
DOSHISHA:TWITTER+INSTAGRAM

	Twi	Ins	sum	s1	s2	s3	s4	result
No1	4994	22417	27411	8561	8871	10620	25957	1st
No2	1310	3513	4823	1833	1876	2178	4353	
No3	3785	11580	15365	5325	5474	6679	13470	
No5	2400	12889	15289	4747	4821	5623	10879	2nd
No6	2471	5832	8303	3485	3514	3923	7215	

TABLE XIX
KANSAI:TWITTER+INSTAGRAM

	Twi	Ins	sum	s1	s2	s3	s4	result
No1	2919	4541	7460	3464	3497	4049	7902	1st
No2	2366	2207	4573	2681	2688	2907	4923	
No3	3694	3182	6876	4164	4178	4491	8942	2nd
No4	2378	3481	5859	2940	2959	3249	6395	
No5	1865	2911	4776	2329	2347	2584	4898	
No6	1662	1902	3564	1955	1963	2128	3696	

TABLE XX
VALUATION:TWITTER+INSTAGRAM

	combination	ordered	winner only	separate
twitter only	3/9	1/9	6/9	7/18
Instagram only	5/9	3/9	6/9	9/18
sum	5/9	3/9	7/9	10/18
s1	5/9	3/9	6/9	9/18
s2	5/9	3/9	6/9	9/18
s3	5/9	3/9	6/9	9/18
s4	4/9	2/9	6/9	8/18

the follower-neighbor graph of their Instagram accounts. All we can easily obtain for their Instagram accounts is the information on the number of followers.

Instead of applying our methods to their Instagram accounts, we developed a method of integrating the information on the number of followers of their Instagram accounts into our prediction. Assuming that the structure of the follower-neighbor graphs of their Twitter accounts and that of their Instagram accounts are similar to some extent, we estimate the score of their Instagram account by the formula below:

$$\text{Insta-score} = \text{Twitter-score} \cdot \frac{\#\text{ Instagram followers}}{\#\text{ Twitter followers}} \cdot \frac{0.106}{0.42}$$

The Insta-score is the score of the user's Instagram account, Twitter-score is the score of the user's Twitter account estimated by one of our methods, # Instagram followers means the number of followers of the user's Instagram account, and # Twitter followers means the number of followers of the user's Twitter account. This formula convert the score of the user's Twitter account into the score of the user's Instagram account by first rescaling the Twitter score up to the size of the Instagram follower count by using the ratio of the number of Twitter followers and the number of Instagram followers,

and also by normalizing the result by multiplying $0.106/0.42$ which is the ratio of the clustering coefficient of Twitter social graph and Instagram social graph. The value 0.106 is the clustering coefficient of Twitter graph reported by Java et al. [14], and 0.42 is the clustering coefficient of Instagram social graph reported by Manikonda et al. [15].

We then sum up the score of the user's Twitter account and the score of the user's Instagram account estimated by the formula above to produce the final score, which integrates information from the Twitter and from the Instagram.

The results of the experiment are shown in Tables XI through XX. In each table, the columns Twi and Ins show the number of Twitter followers and Instagram followers of the user, respectively. The column Total shows the sum of them. These columns corresponds to three baseline method that simply uses these values to predict the winners and the runners-up. The column s1 to s4 shows the prediction based on the sum of the Twitter score by one of our four methods and the Instagram score estimated by our conversion method.

Table XX summarizes the accuracy of the three baseline methods and our four methods. It shows that the simple baseline method that uses the total number of Twitter followers and Instagram followers is the best (including tie) in all the four accuracy metrics. The baseline method that only uses the number of Instagram followers, and our three methods that use variations of the clustering coefficient are tie in two metrics, but inferior to the simple method using the number of Instagram followers in the other two metrics.

Our fourth method, which was the best in the previous experiment, is inferior to the method by the number of Instagram followers in all the four metrics. This result suggests that our method converting the number of Instagram followers into the score by using the Twitter score, in other words, by using the Twitter graph structure, does not work well. It may imply the property of a Twitter account and an Instagram account of the same user are sometimes very different.

Table XX also shows that the method by the number of Instagram followers is better than the method by the number of Twitter followers. It can also be a reason why the method simply using the number of Instagram followers was better than the method that converts it by using the Twitter score.

Nonetheless, when we only use Twitter information, our method outperforms the method that simply use the number of followers, as shown in the previous experiment. Therefore, we expect that our method can outperform the simple method that uses the total number of Twitter followers and Instagram followers if we can retrieve the Instagram social graph and apply our methods to it. Unfortunately, however, Instagram does not allow it for now.

V. Conclusion

In this study, we proposed a method for estimating the real-world popularity of SNS users, which is different from the simple number of followers. Our methods take into account not only the number of followers but also the community structure within the neighbor of the followers. If a user is followed by followers from wider communities, we expect that the user has more global popularity, and more popular in the real world than a user with the similar number of followers only from a smaller local communities.

To validate this hypothesis, we conducted an experiment using data from nine Ms/Mr university competitions. The result shows that the prediction by one of our method achieves higher accuracy than the baseline method that predict the popularity simply by the number of Twitter followers. This result support our hypothesis that a user with followers from wider communities is more popular in the real world.

We also developed a method of integrating the information of the number of Instagram followers into the score by our methods. We convert the number of Instagram followers into a score based on the score of the user's Twitter account and the ratio of the number of user's Twitter followers and the Instagram followers. However, this method was inferior to a simple method that uses the total number of Twitter and Instagram followers. We expect that our methods can achieve higher accuracy if we can retrieve the social graph of Instagram and apply our methods to it.

References

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

[2] D. Imamori and K. Tajima, "Predicting popularity of twitter accounts through the discovery of link-propagating early adopters," in *Proc. of CIKM*, 2016, pp. 639–648.

[3] S. Ohsawa and Y. Matsuo, "Popularity prediction for entities on sns using semantic relations," *Trans. of JSAI: AI*, vol. 29, no. 5, pp. 469–482, 2014.

[4] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring user influence in twitter: The million follower fallacy," in *Proc. of ICWSM*, 2010.

[5] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman, "Influence and passivity in social media," in *Proc. of ECML PKDD*. Springer, 2011, pp. 18–33.

[6] X. Tang and C. C. Yang, "Ranking user influence in healthcare social media," *ACM TIST*, vol. 3, no. 4, pp. 1–21, 2012.

[7] N. Segev, N. Avigdor, and E. Avigdor, "Measuring influence on instagram: a network-oblivious approach," in *Proc. of SIGIR*, 2018, pp. 1009–1012.

[8] J. Saramäki, M. Kivelä, J.-P. Onnela, K. Kaski, and J. Kertész, "Generalizations of the clustering coefficient to weighted complex networks," *Phys. Rev. E*, vol. 75, p. 027105, 2007.

[9] A. Barrat, M. Barthélemy, R. Pastor-Satorras, and A. Vespignani, "The architecture of complex weighted networks," *Proc. of the National Academy of Sciences*, vol. 101, no. 11, pp. 3747–3752, 2004.

[10] J.-P. Onnela, J. Saramäki, J. Kertész, and K. Kaski, "Intensity and coherence of motifs in weighted complex networks," *Phys. Rev. E*, vol. 71, p. 065103, 2005.

[11] G. Fagiolo, "Clustering in complex directed networks," *Phys. Rev. E*, vol. 76, p. 026107, 2007.

[12] K. Berahmand, A. Bouyer, and N. Samadi, "A new centrality measure based on the negative and positive effects of clustering coefficient for identifying influential spreaders in complex networks," *Chaos, Solitons & Fractals*, vol. 110, pp. 41–54, 2018.

[13] S. A. Myers, A. Sharma, P. Gupta, and J. Lin, "Information network or social network? the structure of the twitter follow graph," in *Proc. of WWW Conf. Companion*, 2014, p. 493–498.

[14] A. Java, X. Song, T. Finin, and B. Tseng, "Why we twitter: Understanding microblogging usage and communities," in *Proc. of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, 2007, pp. 56–65.

[15] L. Manikonda, Y. Hu, and S. Kambhampati, "Analyzing user activities, demographics, social network structure and user-generated content on instagram," 2014.