Effects of Ego Networks and Communities on Self-Disclosure in an Online Social Network

Young D. Kwon*, Reza Hadi Mogavi*, Ehsan Ul Haq*, Youngjin Kwon†, Xiaojuan Ma*, and Pan Hui*‡
*Hong Kong University of Science and Engineering, Hong Kong SAR
†Korea Military Academy, Seoul, Republic of Korea
‡University of Helsinki, Helsinki, Finland
Email: {ydkwon, rhadimogavi, euhaq, mxj}@cse.ust.hk kyjchonje@kaist.ac.kr panhui@cs.helsinki.fi

***Abstract*—Understanding how much users disclose personal information in Online Social Networks (OSN) has served various scenarios such as maintaining social relationships and customer segmentation. Prior studies on self-disclosure have relied on surveys or users' direct social networks. These approaches, however, cannot represent the whole population nor consider user dynamics at the community level.**

In this paper, we conduct a quantitative study at different granularities of networks (*ego networks* and *user communities*) to understand users' self-disclosing behaviors better. As our first contribution, we characterize users into three types (open, closed, and moderate) based on the Communication Privacy Management theory and extend the analysis of the self-disclosure of users to a large-scale OSN dataset which could represent the entire network structure. As our second contribution, we show that our proposed features of ego networks and positional and structural properties of communities significantly affect self-disclosing behavior. Based on these insights, we present the possible relation between the propensity of the self-disclosure of users and the sociological theory of structural holes, *i.e.,* users at a bridge position can leverage advantages among distinct groups. To the best of our knowledge, our study provides the first attempt to shed light on the self-disclosure of users using the whole network structure, which paves the way to a better understanding of users' self-disclosing behaviors and their relations with overall network structures.

I. Introduction

Online Social Networks (OSNs) play an essential role as social platforms for millions of users who seek benefits, *e.g.,* people can maintain social relationships [1], [2], improve satisfaction [3], and enjoy more services [4] by disclosing personal information about themselves online. In the literature, self-disclosure means 'the act of revealing personal information to others' [5]. Furthermore, personal information on OSNs published by users benefits various parties. For instance, voluntarily updated and highly identifiable user profiles offer many opportunities for customer segmentation and micro-segmented online advertising [6]. Thus, understanding what factors affect and how much those factors contribute to users' self-disclosure is very important for both users and business intelligence agents.

Several survey-based studies focus on examining why users disclose their personal information online concerning personality traits and contexts of sites [6]–[8]. However, these studies are limited by a small sample size of the survey participants. Some studies applied a semi-supervised or supervised approach to extend the analysis of self-disclosure of users to large-scale data by training their machine learning models on manually-annotated data [9]–[11]. However, previous studies only capture a sampled subgraph where the analysis on the subgraphs may be biased and may not precisely represent all the critical features of the networks [12] as well as neglect user dynamics at the community level which may have a considerable influence on user behaviors [13], [14].

In this paper, we overcome the limitations of the prior works by conducting a quantitative study to better understand users' self-disclosing behaviors in an OSN using a large-scale dataset. We employ the dataset in [15] which consists of 462 million edges among 30 million users. More than 70% of all Google+ users publicly known were collected when the dataset was crawled [15], providing a unique opportunity to capture the comprehensive and unbiased view of network structures on a large scale. We then perform a two-layer network analysis which considers the *ego network* and *user communities* (defined in Section III) to analyze the influence of users' direct social networks and communities, respectively. In this work, we seek to answer the following research questions:

RQ1. What ego network properties can be derived and how much do those features influence the self-disclosure of users?
RQ2. What community properties can be derived and how much do those features influence the self-disclosure of users?
RQ3. To what extent is the self-disclosure of users affected by network properties at the individual and community levels?

Highlights of this work. We are inspired by the Communication Privacy Management theory (known as CPM) to characterize our Google+ users into three types which are *open*, *closed*, and *moderate* [16] (Section III). Open users disclose all their personal information whereas the closed users disclose none; we name the rest of users who lie between as the

ASONAM '19, August 27-30, 2019, Vancouver, Canada
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6868-1/19/08?/$15.00
http://dx.doi.org/10.1145/3341161.3342881

moderate users. We aim to answer RQ1 and RQ2 with extensive analysis on millions of users to differentiate open, moderate, and closed users in their ego networks and communities. To address RQ3, we examine to what extent we can infer the self-disclosure of users by distinguishing their types.

To answer RQ1, we analyze ego networks based on user types to study the propensity of users' self-disclosure (Section IV). In contrast to the prior work [11], our results indicate that the tendency of users to reveal more information positively correlates with their number of friends. Interestingly, moderate users tend to have more dense ego networks than open users (19.7% median increase). This result means that the users with many friends who are not densely connected among them tend to disclose more information. Also, the effective network size which can measure the importance of a user as a bridge [17] is significantly higher for open and moderate users than for closed users. It seems that users are more likely to reveal information when they are in bridge positions where they can utilize positional advantages. Hence, we present the potential relations between the propensity of self-disclosure of users and the sociological theory of structural holes (*i.e.,* users in a bridge position can leverage the advantages by connecting distinct groups from the networks) [17]).

To answer RQ2, we investigate users' self-disclosure in the contexts of communities with the intuition that a community can influence its users' behaviors [13], [14] (Section V). We discover that users' self-disclosing behaviors manifest themselves differently based on the positional and structural properties of communities. For example, users located in a critical position (*i.e., bridge position*) within a community tend to disclose more personal information, which can also be explained by the structural holes theory. Besides, the average amount of personal information revealed in a community significantly differs according to the structural properties of communities such as the average density of community and community size.

For answering RQ3, we formulate a classification task to distinguish user types using all network properties we study given only a network structure (Section VI). We then identify the importance of the proposed features from the learned models using our features. This result further validates our insight that structural holes may justify high levels of the self-disclosure of users whose position is at the bridge in a network. In summary, the main contributions of this work are:

1) We extend the analysis of users' self-disclosing behaviors to the large-scale dataset by characterizing them into three user types based on the online privacy theory, CPM.
2) We study self-disclosure of users concerning two different levels of granularity, ego networks and user communities, and present the possible explanation for users' self-disclosing behaviors using the sociological theory of structural holes.
3) We explore the possibility of inferring the self-disclosure levels of users given that we can only access the structural information of an OSN as well as confirm the importance of the features relevant to the structural holes theory.

The rest of the paper is organized as follows. Section II summarizes the related work and Section VII concludes the paper and presents future directions.

II. Related Work

To begin, we review the literature on network properties and self-disclosure.

Network properties. Many works have focused on studying the network properties of communities in OSNs to understand and model user behavior, user engagement in communities, and community evolution [18], [19]. Network properties like clustering coefficients show the network evolution [20], [21]. Graph centrality and the betweenness property are used along with closeness to predict influential users. Many studies have used network properties to understand user behaviors [22]–[26], however, no prominent work has related such network properties to users' self-disclosing behaviors.

Self-disclosure. Researchers have investigated self-disclosing behaviors of users using survey data [6]–[8] or manually-annotated datasets [9]–[11], [27]. With survey datasets, Krasnova et al. found users are motivated to disclose their personal information by the convenience of developing and maintaining social relationships through OSNs [6]. Users' behaviors when disclosing personal information were studied in terms of their personalities [8] and demographics with different contexts such as private and business OSNs [7]. With manually-annotated datasets, past works largely focused on linguistic features. Bak et al. developed a topic model to classify users on Twitter according to three self-disclosure levels: High, Low, and No disclosure [9]. Wang et al. studied self-disclosure with linguistic features and ego network properties [11]. Several studies focused on a specific type of self-disclosure such as depression [28], [29].

We, however, perform the analysis of users' self-disclosing behaviors with a large-scale dataset as well as explore novel relations between the self-disclosure of users and two levels of networks: ego networks and user communities.

III. Preliminaries

Dataset. We employ Google+ dataset collected by Gong et al. [15]. It spans over three months from July 6th to October 11th, 2011 and consists of 30 million users and 462 million edges. The authors collected more than 70% of the whole user base publicly known when the dataset was crawled by [15], which provides a unique opportunity to capture the comprehensive and unbiased view of network structures on a large scale. In Google+, users are allowed to customize their profiles. In the dataset, there are four optional attribute types where users can choose whether to provide their personal information as follows: *City*, *School*, *Major*, and *Employer*. As we see in Figure 1, the dataset provides a social graph structure and optional attributes for each user.

Ego networks, communities, and structural holes. In contrast to a global social network, an ego network is a personal social network which consists of two components: an *ego* and *alters* [30]. The ego is a single user and alters are all connected

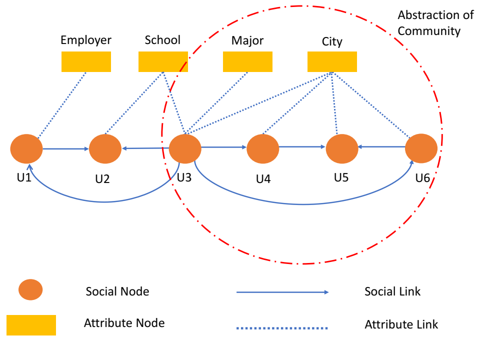


Figure 1: Illustration of an OSN with an abstract community. Social nodes have connections among themselves and can be linked with attribute nodes.

neighbors to the user. For example, the ego network of user u consists of all the users that she follows and all the users that are following her (*i.e.,* all the incoming and outgoing edges of u). In the literature, connected social nodes in a network construct communities [31]. The *Structural holes* [17] appear when distinct groups of users are connected through user u and become disconnected if user u is not included, in which user u plays a bridge role connecting those distinct groups of users. The structural holes allow users as bridges to leverage positional advantages between separate groups in terms of information benefits [32].

A. Characterization of Users

A crucial contribution in this research is to characterize users according to the extent they are willing to *disclose* their personal information in OSNs. Since the privacy concerns of users are highly diverse and subjective, prior work has relied mainly on running surveys or manual annotations. Although these methods work favorably for a small number of samples, they suffer two problems in dealing with large-scale datasets: (1) a survey-based method is difficult to generalize to the whole population since the number of samples is limited; (2) a manual annotation method suffers from the lack of ground truth and disagreement among annotators. To tackle the drawbacks mentioned above, we use a data-driven approach on a large-scale dataset with more than 70% of the entire user base to study users' privacy concerns. Furthermore, in order to better highlight the differences, we are inspired by a well-known online privacy theory called CPM to categorize Google+ users into three groups and track their privacy-related features [16]. Google+ suggests its users fill in four optional personal information fields. We name users who have filled in all four fields of optional information as *open users* since they fully disclose all their information. Conversely, we name users who have not provided any information as *closed users* since they do not reveal any non-mandatory information in their profiles. We refer to the rest of the users as *moderate users*.

We use the initial stage of the dataset which can represent the entire OSN but not too large to compute a computationally expensive network feature (*e.g.,* betweenness centrality). Hence, we identify 3.78M closed users, 786K moderate users, and 123K open users. The ratios of the number of closed, moderate, and open users are 80.6%, 16.8%, and 2.6%, respectively.

B. Notation

In a social network G, nodes are users and edges represent relationships between users. Users and attributes in G are called *social nodes* and *attribute nodes*, respectively. While social nodes are denoted as set V_s, attribute nodes are denoted as set V_a. In addition, *social links* and *attribute links* indicate links between social nodes, and links between social nodes and attribute nodes, respectively. Set E_s indicate social links and set E_a indicate attribute links. Lastly, the set (V_s, V_a, E_s, E_a) represents an OSN.

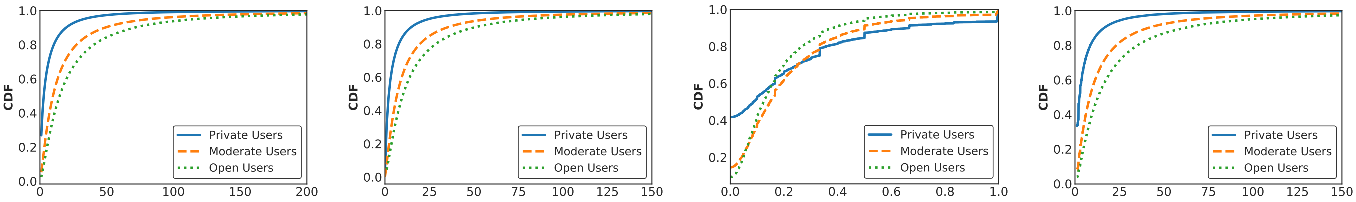
IV. Self-Disclosure in Ego Networks

The primary goal of this section is to study the self-disclosure behaviors of users at an individual level in their ego networks according to their user types. Prior work [11] showed that the number of friends negatively correlates with self-disclosure in the context of a private OSN such as Facebook since having more friends implies that users have less control of the revelation of their information online. However, users who use the Google+ OSN associate their jobs while interacting on the platform [33], which makes us conjecture that users in a more public OSN would behave differently. To answer **RQ1**, we first investigate the differences among closed, moderate, and open users using degree and the clustering coefficient (CC), as studied in [11], as well as effective network size. After that, we explain our new observations based on the sociological theory of structural holes [32]. For hypothesis tests, we conduct a Kruskal-Wallis Test (KW Test) to statistically validate distinctive patterns among different groups for each feature and Mann-Whitney's U Test (U Test) for the post-hoc test. In particular, we hypothesize as follows:

> ***H1.1:H1.3*** *There exist significant differences in ego network properties among three user types, and open users have significantly higher (H1.1: degrees, H1.2: CC, H1.3: effective network sizes) than other user types.*

Degree (H1.1). The number of unique connections a user has to other nodes in a graph is called a degree. The degree is often referred to as the *network size* in ego networks. As shown in Figure 2a, the results show that user types have a significant effect on the medians of degrees (KW Test: $\chi^2(2) = 505K, p < 0.001$). Open users have significantly higher degrees than moderate (U Test: $Z = -107.3, p < 0.001, r = 0.11$) and closed users (U Test: $Z = -359.5, p < 0.001, r = 0.18$); the median degree of open users is 50.0% and 400.0% greater than the median of moderate and closed users, respectively. Moderate users tend to have higher degrees than closed users (233.3% median increase) and the difference is significant (U Test: $Z = -638.6, p < 0.001, r = 0.30$). In addition, as in 2b, indegrees show similar patterns to degrees (This also holds for outdegrees). This result is in contrast to the previous finding that a degree is negatively associated with self-disclosure [11].

Clustering coefficient (H1.2). We examine how CC affects a user's propensity of self-disclosure online. CC measures how



(a) Degree (b) InDegree (c) Clustering Coefficient (d) Effective Network Size

Figure 2: Structural differences in ego networks of closed users, moderate users, and open users. The x-axis represents feature values based on an individual's ego network and the y-axis represents the percentage.

much a user's neighbors know each other. CC is also referred to the *density* in ego networks. In our graph, the CC for a user u is given by:

$$CC(u) = \begin{cases} \frac{\{|(s,t);(s,t)\in E_s|\}}{d_u(d_u-1)} & \text{if } d_u > 1 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

s and t are u's neighbors and d_u is u's degree and E_s covers social links of u's ego network. As shown in Figure 2c, the results show that user types have a significant effect on the medians of CC (KW Test: $\chi^2(2) = 59K, p < 0.001$). Closed users have a significantly smaller CC than moderate users (U Test: $Z = -235.5, p < 0.001, r = 0.11$); the median CC of closed users is 41.7% smaller than the median of moderate users. While the effect size is small, closed users have a smaller CC than open users (U Test: $Z = -82.8, p < 0.001, r = 0.04$). However, between open and moderate users, moderate users tend to have a higher CC than open users (19.7% median increase) (U Test: $Z = 39.9, p < 0.001, r = 0.04$). These results partially follow the previous finding that CC positively correlates with self-disclosure [11].

Effective network size (H1.3). We study the relationship between the effective network size and the propensity of self-disclosure. The effective network size refers to the ego network size of distinct groups of neighbors in terms of information benefits [17]. These distinct groups of users who connect through user u become disconnected if user u is ignored. Hence, the effective network size indicates how much user u plays a bridge role in her ego network. According to the structural holes theory, when neighbors of user u from several groups are densely interconnected, only one neighbor from each distinct group is considered non-redundant in terms of information benefits. The remainders are redundant since the one neighbor in each distinct group can still provide the same information that other redundant neighbors provide to user u [17]. Hence, the effective size of u's ego network is defined as follows:

$$\text{Effective network size} = d_u - \frac{\{|(s,t);(s,t)\in E_s|\}}{d_u} \quad (2)$$

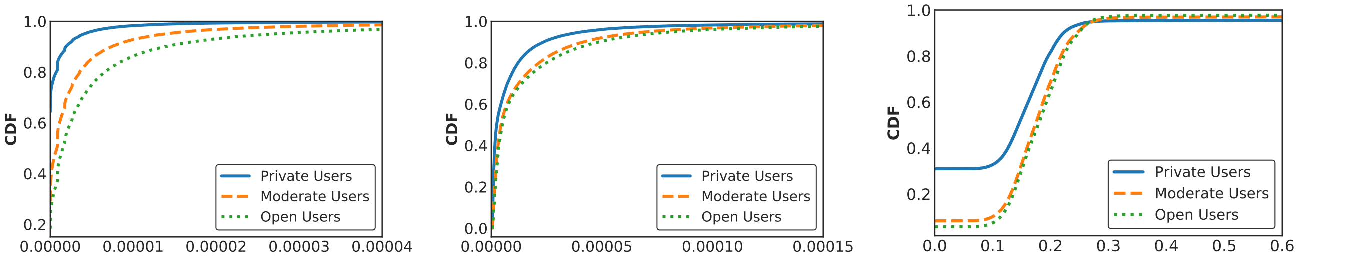
where s and t are u's neighbors and d_u is u's degree, and E_s covers social links of u's ego network. The effective network size varies from 1 where all neighbors are redundant to d_u where all neighbors are non-redundant. Figure 2d shows that the larger the effective network size, the more open the users are. The differences in user types on the median of the effective network size are significant (KW Test: $\chi^2(2) = 510K, p < 0.001$). We show that open users have significantly greater effective network size than moderate (U Test: $Z = -113.8, p < 0.001, r = 0.11$) and closed users (U Test: $Z = -366.8, p < 0.001, r = 0.18$); the median effective network size of open users is 54.4% and 414.7% greater than the median of moderate and closed users, respectively. Moderate users tend to have a higher effective network size than closed users (233.3% median increase) and the difference is significant (U Test: $Z = -638.7, p < 0.001, r = 0.30$).

Discussion. We believe that the characteristics of the Google+ platform and the theory of social structural holes [17] can explain our new observations of the patterns of users' self-disclosing behaviors. First, users who use Google+ OSN do not dissociate from their jobs while interacting on the platform [33]. Thus, while users have a wider audience (*i.e.,* more friends), they may not be discouraged to update their personal information from being concerned about disclosure of personal information as in private OSN (*e.g.,* Facebook [11]). Second, based on our observations of the CC and effective network size, the structural holes may play an important role in the self-disclosure of users. For example, when most of the neighbors of the users are densely connected, the users may not be motivated to reveal more information to utilize the network because their neighbors cannot provide new information benefits [32]. Moreover, the effect of the effective network size on user types implies the importance of the bridging role of a user.

Overall, we confirm and extend the prior work [11] on the large-scale dataset in the more public context of OSN, and show that the structural holes may play an essential role in the self-disclosure of users.

V. Self-Disclosure in Communities

We have observed that features in an individual level network affect users' propensity to disclose information in an OSN. We now turn our attention to the community side since the online and offline behaviors of users can be highly influenced by various properties of communities [13], [14]. We start by identifying communities from the dataset. After that, we explore how community features such as a user's positional properties



(a) Betweenness Centrality

(b) PageRank

(c) Closeness Centrality

Figure 3: Comparison of three user groups based on different positional properties in the contexts of communities.

in the context of communities and the structural identity of the community affect users' self-disclosure to answer **RQ2**.

A. Identification of Communities

Communities form with highly interconnected users. To evaluate the network structure with communities, *modularity* is a widely used metric to explain how well communities cluster in a network structure [31]. Larger modularity means the network can be clustered into communities effectively. The modularity is defined as follows:

$$Modularity = \frac{1}{2m}\sum_{ij}(A_{ij} - \frac{k_i k_j}{2m})\delta(c_i, c_j) \tag{3}$$

where m refers to the total number of edges, A refers to the adjacency matrix with the values of 1 if node i and j are connected or of 0 otherwise, k_i means the degree of node i, and $\delta(c_i, c_j) = 1$ if node i and j belong to the same community or $\delta(c_i, c_j) = 0$ otherwise. To take into account properties of a community, we use the Louvain method [34] to identify communities and further demonstrate the community structure of the employed dataset. The Louvain method is a bottom-up approach which iteratively groups nodes and communities until the gain of modularity falls below a threshold of λ. Since the λ parameter is a vital tuning parameter for the Louvain method [35], we verify the quality of the detected community structure by running the Louvain method with different threshold λ values ranging from 10^{-8} to 0.4 which consistently show high modularity values of above 0.7 [36]. We adopt λ of 10^{-7} for our further study since it shows high modularity of 0.77. Note that the network structure with modularity $\geq$ 0.3 indicates the corresponding network has a well-formulated community structure [37]. This result verifies that our detected communities are well organized as a community structure with a given network G. As a result of community detection, we identify a total of 4055 communities in the dataset.

B. Positional Properties in Communities

After identifying communities, we are interested in whether the relative position of users within the community affect users' self-disclosing behaviors. For the positional properties in communities that users belong to, we adopt three widely used node centrality measures: (1) betweenness centrality (BC), (2) PageRank centrality (PR), and (3) closeness centrality [38]. More specifically, we predict as follows:

> ***H2.1:H2.3*** *Among three user types, there exist significant differences in positional properties in communities, and open users have significantly higher (H2.1: BC, H2.2: PR, H2.3: closeness centrality) than other user types.*

Betweenness centrality (H2.1). BC is a metric that quantifies how central a node's role is in connecting other nodes. Hence, we use BC [39] to examine the effects of a *bridging* role in a community on the self-disclosure of users. BC measures the sum of the fraction of the all-pairs shortest paths that pass through user u and is given by:

$$BC(u) = \sum_{s,t \in V_s, s \neq u \neq t} \frac{\sigma(s,t|u)}{\sigma(s,t)} \tag{4}$$

where $\sigma(s,t)$ is the number of shortest paths between user s and user t, and $\sigma(s,t|u)$ is the number of shortest paths between user s and user t passing through user u. Figure 3a shows that users tend to disclose more personal information as they have a higher BC. The differences among user types are significant (KW Test: $\chi^2(2) = 515K, p < 0.001$). The result shows that open users have significantly higher BC than moderate users (U Test: $p < 0.001$, 87.5% median increase).

PageRank centrality (H2.2). We study PR [40] which represents a user's importance in a given community. A larger PR value indicates that the corresponding node is more important and influential [36]. The PR index for a user in a community of size N is given by:

$$PR(u) = \frac{1-\alpha}{N} + \alpha \sum_{v \in M(u)} \frac{PR(v)}{d_v^{out}} \tag{5}$$

where α is a damping factor set as 0.85, and $M(u)$ is the set of nodes that link to u, and d_v^{out} is the number of outgoing links from node v. As in Figure 3b, the more important the user is, the more information the other users provide. The differences among user types are significant (KW Test: $\chi^2(2) = 122K, p < 0.001$). However, the post-hoc test reveals that the effect size of PR is relatively small between moderate and open users (U Test: $Z = -32.0, p < 0.001, r = 0.03$).

Closeness centrality (H2.3). We investigate the closeness centrality which measures how close a user is to all other users

Table I: Statistics for structural properties of communities.

Features	Min	Q1	Median	Q3	Max
Community size	2	3	4	6	581403
NACC	0.00	0.00	0.00	0.34	1.00
Avg. degree	1.00	2.00	2.50	3.50	45.34
Avg. SP	0.05	0.53	0.96	1.24	6.48
Diameter	1.00	1.00	2.00	3.00	41.00

in a community. In our study, we define the closeness centrality for a user u in a community of size N as follows:

$$Closeness(u) = \frac{N-1}{\sum_{v \in V_s^k} SP(v,u)} \tag{6}$$

where $SP(v,u)$ is the shortest-path distance between a user v and user u. V_s^k denotes the set of all users in the community k. Figure 3c indicates that the closer a user is to all other users in a community, the more likely the user is to reveal her information. There exist significant differences among user groups (KW Test: $\chi^2(2) = 192K, p < 0.001$). Note that over 30% of closed users have 0 closeness centrality because they only have outgoing links and thus other users cannot reach them. Likewise PR, closeness centrality has a small effect size between moderate and open users (U Test: $Z = -29.8, p < 0.001, r = 0.03$).

Discussion. From the results of statistical analysis, we derive several findings as follows: (1) overall, users who are in an important position in a community disclose more information; (2) in particular, being positioned as a bridge in a community (having a high BC) is one of the major factors for the self-disclosure of users as we use BC as a TOP 3 important feature for building the machine learning model in Section VI; (3) we further confirms the importance of structural holes as we have discussed in Section IV.

C. Structural Properties of Communities

We are inspired by [35] and turn our attention to whether the structural characteristics of the community itself affects how much users reveal their personal information (*i.e.,* the number of attribute types). For that, we first derive several features that capture the structural properties of communities as follows: (1) community size, (2) network average clustering coefficient (NACC), (3) average degree, and (4) distance measures. To examine in detail how users in a community with different structural properties behave, we further group users from communities according to the value ranges defined for each structural property. For community size, following [35], we categorize community sizes into four groups and denote them as follows. [x, y] represents communities sized between x and y (*e.g.,* [Min,10], [10,1K], [1K,10K], and [10K+]). For NACC, we divide users into three groups since the percentile statistics of NACC are 0 from the minimum to the median. Then, for the rest of the structural properties of communities, the value ranges are categorized according to 25th percentile, median, and 75th percentile. Table I reports the descriptive statistics for structural properties. In this section, we use the average number of attribute types of a community (*i.e.,* overall self-disclosure levels of users in the community) as a dependent variable for hypothesis tests. We then predict as follows:

H2.4:H2.7 *There exist significant differences in the average number of attribute types of communities among four groups of different (H2.4: community sizes, H2.5: NACC, H2.6: average degrees, H2.7: distance measures).*

Community size (H2.4). We examine how users reveal their personal information regarding community size. The community size is defined as the number of all users in a community. The result shows that users' self-disclosing behaviors significantly depend on the size of the community (KW Test: $\chi^2(3) = 127.35, p < 0.001$). However, U Test result ($p = 0.775$) does not show any significant effect of community size between medium-sized communities (size of [1K,10K]) and big communities (size of 10K+), while medium-sized communities' average number of attribute types are slightly higher ($Median = 0.424$) than big communities ($Median = 0.416$). This result is partially in accordance with the previous finding that a relatively small platform appears to foster a more congenial atmosphere [41], which may lead to more self-disclosure of the personal information of users.

Network average clustering coefficient (H2.5). We study how users reveal personal information in relation to NACC. NACC is the average CC of all users in an OSN, *i.e.,* the average density of an OSN. In our study, we consider the NACC of each community k and NACC is defined as follows:

$$NACC(k) = \frac{1}{|V_s^k|} \sum_{u \in V_s^k} CC(u) \tag{7}$$

where $|V_s^k|$ is the total number of users in the community k. We discover that users' self-disclosure largely depends on and is proportional to the NACC of the community (KW Test: $\chi^2(2) = 229.40, p < 0.001$). Like community size, NACC does not have a significant effect on self-disclosure between communities with a medium NACC and communities with a high NACC (U Test: $p = 0.146$).

Average degree (H2.6). We investigate the way users disclose personal information in relation to the average degrees of all users in a community. The result shows that users tend to provide more personal information as the average degree of a community gets bigger from min to max (KW Test: $\chi^2(3) = 360.31, p < 0.001$).

Distance measures (H2.7). We finally examine how users disclose their personal information concerning two distance-related features. Eccentricity, *i.e.*, the maximum distance among the shortest path lengths from a user u to all other users in community k, determines distance measures of a community. $Eccentricity(u)$ is equal to $argmax_{v \in V_s^k} SP(u,v)$ where $SP(u,v)$ is the shortest path length from node u to node v. The average shortest path length and diameter of community k are defined as the average and maximum eccentricity of community k, respectively. We find that users in a community with higher values of the average shortest path length reveal more information (KW Test: $\chi^2(3) = 228.67, p < 0.001$).

Table II: Performance of models learned with different features on distinguishing user types. The results show improvements from incrementally adding our proposed features to strong indicators.

Feature Sets	Precision	Recall	F1
Degree & CC [11]	0.655	0.615	0.633
+ Ego Network Properties	0.656	0.745	0.697
+ Positional Properties	0.662	0.760	0.707
+ Structural Properties (Full)	0.666	0.761	0.710

Diameter also shows a similar pattern to the average shortest path length (KW Test: $\chi^2(3) = 325.70, p < 0.001$).

Discussion. Overall, we find that all structural properties of communities have a significant effect on the average number of attribute types of a community (*i.e.,* overall self-disclosure levels of users in the community).

VI. Inferring Self-Disclosure

Having established the features from ego networks and communities, we investigate to what extent the self-disclosure of users is affected by all of the network properties examined thus far. In this section, to answer **RQ3**, we formulate a classification task. This step allows us (1) to identify the importance of the network properties, (2) to further validate a previously described insight, the importance of the features relevant to the structural holes (*i.e.,* Effective network size and BC), and (3) to explore the possibility of inferring the self-disclosure of users given that we can only access the structural information of OSN.

A. Experimental Setup

Given a social network of V_s and E_s, we formulate a classification task to distinguish moderate and open users from closed users. Since the proportions of user types are highly imbalanced, we randomly select 3,000 users from each user group by using a random under-sampling method [42]. To overcome the potential bias in our sampled datasets and obtain the generalizability of our results, we conduct the experiments over 100 randomly sampled datasets. We determine 90% of users as training/evaluation sets and the remaining 10% of users as a test set. Then we use precision, recall, and the F1-score as evaluation metrics with moderate and open users as the target class.

Benchmark and proposed features. Prior work on self-disclosure found network features, degree, and CC, are powerful indicators to infer users' self-disclosure using a regression model [11]. We first build a strong benchmark using these indicators [1]. After that, we further propose features based on the properties of ego networks and communities we have studied in the previous sections. To study the effects of our proposed features on the classification task, we build models for inferring self-disclosure of users by using our proposed features as follows.

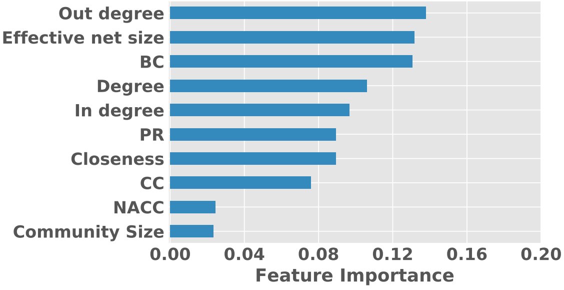


Figure 4: Feature importance for distinguishing user types.

1) **Ego network properties:** Degree, in/out degree, CC, and effective network size (Section IV).
2) **Positional properties in communities:** BC, PR, and closeness centrality (Section V-B).
3) **Structural properties of communities:** Community size, NACC, average degree, average shortest path length, and diameter (Section V-C).

B. Overall Performance

We first train four different classifiers such as Logistic Regression, Multiple-Layer Perceptron, Decision Tree, and Random Forest (RF). Among them, we use RF for further comparison between different features since RF shows the best performance and enables us to calculate the importance of each feature. Note that we do not adopt deep neural network models in our study because our primary goal is to have a better understanding of our derived features and their relative importance. Table II shows that the full model using all proposed features performs the best, and significantly improves the performance of the benchmark by 12% in the F1-score (Wilcoxon signed-rank test: $p < 0.001$).

C. Understanding Feature Importance

We finally discuss the feature importance of the learned RF with all features. Figure 4 lists the top 10 important features. We calculate the Gini importance of each feature in the learned RF from the classification task. It suggests that ego network properties and features derived from the contexts of communities are important for distinguishing user types. The BC and effective network size are two of the top 3 most important features, which is consistent with our finding that the users' roles as bridges in a community are important to self-disclosing behaviors.

VII. Conclusion

In this paper, we studied the self-disclosure of users based on their ego networks and communities. We first characterized the self-disclosure of users into three types following online privacy theory called CPM to extend the survey-based analysis of self-disclosure to the large-scale dataset. After that, we examined how users' self-disclosing behaviors are associated with network structures in two levels of granularity: ego networks and user communities. Interestingly, we observed that self-disclosing behaviors could be associated with the

[1] We do not include other studied features such as gender, age, number of logins, and tie strength since these features are not available in the dataset.

sociological theory of structural holes since users are more likely to disclose personal information when they can utilize positional advantages by playing bridging roles from their networks. This work established the potential to incorporate different network structures into self-disclosure analysis.

There are several directions worthwhile investigating as future work. First, we want to verify our results with other data sources. Second, we plan to explore the possibility of predicting the future status of the self-disclosure of users dynamically, which can provide much information to various business agents. Finally, we want to investigate the causality relation between the self-disclosure and the user's position in a network.

Acknowledgement

This research has been supported in part by project 16214817 from the Research Grants Council of Hong Kong and the 5GEAR project from the Academy of Finland.

References

[1] N. L. Collins and L. C. Miller, "Self-disclosure and liking: a meta-analytic review," *Psychological Bulletin*, vol. 116, no. 3, pp. 457–475, Nov. 1994.

[2] N. Park, B. Jin, and S.-A. Annie Jin, "Effects of self-disclosure on relational intimacy in Facebook," *Computers in Human Behavior*, vol. 27, no. 5, pp. 1974–1983, Sep. 2011.

[3] D. L. Oswald, E. M. Clark, and C. M. Kelly, "Friendship Maintenance: An Analysis of Individual and Dyad Behaviors," *Journal of Social and Clinical Psychology*, vol. 23, no. 3, pp. 413–441, Jun. 2004.

[4] S. Sprecher, S. Treger, and J. D. Wondra, "Effects of self-disclosure role on liking, closeness, and other impressions in get-acquainted interactions," *Journal of Social and Personal Relationships*, vol. 30, no. 4, pp. 497–514, Jun. 2013.

[5] R. L. Archer, "Self-disclosure the self in social psychology (pp. 183-205)," 1980.

[6] H. Krasnova, S. Spiekermann, K. Koroleva, and T. Hildebrand, "Online social networks: why we disclose," *Journal of Information Technology*, vol. 25, no. 2, pp. 109–125, Jun. 2010.

[7] A. K. Schaar, A. C. Valdez, and M. Ziefle, "The Impact of User Diversity on the Willingness to Disclose Personal Information in Social Network Services," ser. Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, 2013, pp. 174–193.

[8] E. T. Loiacono, "Self-Disclosure Behavior on Social Networking Web Sites," *International Journal of Electronic Commerce*, vol. 19, no. 2, pp. 66–94, Apr. 2015.

[9] J. Bak, C.-Y. Lin, and A. H. Oh, "Self-disclosure topic model for classifying and analyzing Twitter conversations," ser. EMNLP, 2014.

[10] S. Balani and M. De Choudhury, "Detecting and Characterizing Mental Health Related Self-Disclosure in Social Media," ser. CHI EA '15. New York, NY, USA: ACM, 2015, pp. 1373–1378.

[11] Y.-C. Wang, M. Burke, and R. Kraut, "Modeling Self-Disclosure in Social Networking Sites," ser. CSCW '16. New York, NY, USA: ACM, 2016, pp. 74–85.

[12] T. Wang, Y. Chen, Z. Zhang, T. Xu, L. Jin, P. Hui, B. Deng, and X. Li, "Understanding Graph Sampling Algorithms for Social Network Analysis," in *2011 31st International Conference on Distributed Computing Systems Workshops*, Jun. 2011, pp. 123–128.

[13] W. W. Zachary, "An Information Flow Model for Conflict and Fission in Small Groups," *Journal of Anthropological Research*, vol. 33, no. 4, pp. 452–473, Dec. 1977.

[14] J. Zhang, W. L. Hamilton, C. Danescu-Niculescu-Mizil, D. Jurafsky, and J. Leskovec, "Community Identity and User Engagement in a Multi-Community Landscape," vol. 2017, 2017, pp. 377–386.

[15] N. Z. Gong, W. Xu, L. Huang, P. Mittal, E. Stefanov, V. Sekar, and D. Song, "Evolution of Social-attribute Networks: Measurements, Modeling, and Implications Using Google+," ser. IMC '12. ACM, 2012, pp. 131–144.

[16] S. Trepte and L. Reinecke, *Privacy online: Perspectives on privacy and self-disclosure in the social web*. Springer Science & Business Media, 2011.

[17] R. S. Burt, "Structural Holes: The Social Structure of Competition," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 1496205, 1992.

[18] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, "Characterizing User Behavior in Online Social Networks," ser. IMC '09. New York, NY, USA: ACM, 2009, pp. 49–62.

[19] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3, pp. 75–174, Feb. 2010.

[20] X.-G. Wang, "A network evolution model based on community structure," *Neurocomputing*, vol. 168, pp. 1037–1043, Nov. 2015.

[21] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins, "Microscopic Evolution of Social Networks," ser. KDD '08. ACM, 2008, pp. 462–470.

[22] B. Ngonmang, E. Viennet, and M. Tchuente, "Churn Prediction in a Real Online Social Network Using Local CommunIty Analysis," ser. ASONAM '12, Washington, DC, USA, 2012, pp. 282–288.

[23] R. J. Oentaryo, E.-P. Lim, D. Lo, F. Zhu, and P. K. Prasetyo, "Collective Churn Prediction in Social Network," ser. ASONAM '12, Washington, DC, USA, 2012, pp. 210–214.

[24] N. Phan, D. Dou, B. Piniewski, and D. Kil, "Social restricted Boltzmann Machine: Human behavior prediction in health social networks," Aug. 2015, pp. 424–431.

[25] P. Devineni, E. E. Papalexakis, D. Koutra, A. S. Doğruöz, and M. Faloutsos, "One Size Does Not Fit All: Profiling Personalized Time-Evolving User Behaviors," ser. ASONAM '17. New York, NY, USA: ACM, 2017, pp. 331–340, event-place: Sydney, Australia.

[26] Y. Ren, V. Cedeno-Mieles, Z. Hu, X. Deng, A. Adiga, C. Barrett, S. Ekanayake, B. J. Goode, G. Korkmaz, C. J. Kuhlman, D. Machi, M. V. Marathe, N. Ramakrishnan, S. S. Ravi, P. Sarat, N. Selt, N. Contractor, J. Epstein, and M. W. Macy, "Generative Modeling of Human Behavior and Social Interactions Using Abductive Analysis," ser. ASONAM '18, Aug. 2018, pp. 413–420.

[27] K. Jaidka, S. C. Guntuku, and L. H. Ungar, "Facebook versus Twitter: Differences in Self-Disclosure and Trait Prediction," ser. ICWSM '18, Jun. 2018.

[28] Nazanin Andalibi, Pinar Öztürk, and Andrea Forte, "Sensitive Self-disclosures, Responses, and Social Support on Instagram: The Case of #Depression," ser. CSCW, 2017.

[29] D. Yang, Z. Yao, and R. Kraut, "Self-Disclosure and Channel Difference in Online Health Support Groups," ser. ICWSM '17, May 2017.

[30] V. Arnaboldi, M. Conti, A. Passarella, and R. I. M. Dunbar, "Online Social Networks and information diffusion: The role of ego networks," *Online Social Networks and Media*, vol. 1, pp. 44–55, Jun. 2017.

[31] M. E. J. Newman, "Analysis of weighted networks," *Physical Review E*, vol. 70, no. 5, Nov. 2004.

[32] R. Burt, "Structural Holes and Good Ideas," *American Journal of Sociology*, vol. 110, no. 2, pp. 349–399, 2004.

[33] E. Cunha, G. Magno, M. A. Gonçalves, C. Cambraia, and V. Almeida, "How You Post is Who You Are: Characterizing Google+ Status Updates Across Social Groups," ser. HT '14. ACM, 2014, pp. 212–217.

[34] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.

[35] X. Zhao, A. Sala, C. Wilson, X. Wang, S. Gaito, H. Zheng, and B. Y. Zhao, "Multi-scale Dynamics in a Massive Online Social Network," ser. IMC '12. New York, NY, USA: ACM, 2012, pp. 171–184.

[36] Y. Chen, J. Hu, H. Zhao, Y. Xiao, and P. Hui, "Measurement and Analysis of the Swarm Social Network With Tens of Millions of Nodes," *IEEE Access*, vol. 6, pp. 4547–4559, 2018.

[37] H. Kwak, Y. Choi, Y.-H. Eom, H. Jeong, and S. Moon, "Mining Communities in Networks: A Solution for Consistency and Its Evaluation," ser. IMC '09. New York, NY, USA: ACM, 2009, pp. 301–314.

[38] R. Zafarani, M. A. Abbasi, and H. Liu, *Social Media Mining: An Introduction*. Cambridge University Press, Apr. 2014.

[39] U. Brandes, "A faster algorithm for betweenness centrality," *The Journal of Mathematical Sociology*, vol. 25, no. 2, pp. 163–177, Jun. 2001.

[40] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Technical Report 1999-66, November 1999.

[41] E. Newell, D. Jurgens, H. Saleem, H. Vala, J. Sassine, C. Armstrong, and D. Ruths, "User migration in online social networks: A case study on reddit during a period of community unrest," 2016.

[42] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 20–29, Jun. 2004.