Extremely Low Reciprocity and Strong Homophily in the World Largest MSM Social Network

Mengsi Cai, Ge Huang, Mirjam E. Kretzschmar, Xiaohong Chen and Xin Lu

Abstract—Traditional survey-based methods are limited in the sample size and inference ability for the study of men who have sex with men (MSM), one of the most vulnerable groups at increased risk of HIV. Internet data, on the other hand, have provided publicly accessible information about such groups at unprecedented scale and resolution. Here we present statistics for the user demographics characteristics in the world's largest MSM geosocial networking application, Blued, and analyze the social network structure with 11,408,872 nodes and 838,910,078 edges extracted from user-following relationships on Blued. Network features, such as degree distribution, reciprocity, degree assortativity, homophily, and community are studied. We find that, in contrast to earlier analyses on social networks of general populations, the MSM social network is disassortative and shows extremely low reciprocity. Users in their twenties are excessively followed by users from all age groups, and network homophily for age and country are strong.

Index Terms—Social network; MSM; Social networking application; Network analysis; Blued

1 Introduction

Men who have sex with men (MSM) are more likely to be HIV-infected than general populations, and continue to be a major target population for HIV prevention and intervention [1]. Currently, MSM-related studies have focused on HIV-related issues [2],[3],[4], sexual behaviors [5],[6], social support [7],[8], stigma [9], and mental health [10], most of which are conducted by survey-based methods [11], such as snowball sampling [12], respondent-driven sampling (RDS) [13],[14], or partner notification [15]. These survey-based methods have to some extent improved the accessibility to MSM populations, but have shown underperformance in terms of sample size, convenience, and accessibility [2]. Regarding the hard-to-access properties of MSM populations [2],[16] , an effective and reliable strategy for recruiting MSM populations with large sample sizes is required.

With the development of Internet technology, geosocial networking (GSN) applications, such as Blued, Grindr, and Hornet, have developed as popular platforms for MSM to socialize and seek partners since 2009 (see Table 1). The common usage of MSM-targeted GSN apps has led to the accumulation of a large amount of social networking data, including user profiles, textual content, online behaviors, and social interactions, which offer a new opportunity for understanding MSM populations with unprecedented data volume and richness of information. For example, the world's largest MSM GSN app named Blued, originating from China, where same-sex marriage is not legally recognized, is now popular with over 40 million MSM from over 200 countries [17]. With the high acceptability of GSN apps among MSM, using GSN apps has become an effective way to reach MSM communities and disseminate health interventions [18].

In recent years, many studies have been conducted to examine the association between the use of GSN apps and MSM sexual health. For example, Bien et al. organized a cross-sectional online survey on MSM, which showed that app users are more likely to be younger, better educated, and more likely to report multiple recent sex partners and HIV testing than non-app users [19]. Using venue-based sampling, Phillips et al. reported that the majority of MSM used GSN applications to find sexual partners in the past year, and nearly one-quarter of MSM had sex with a man they met using a GSN app in the prior year [20]. Through a computer-assisted self-interview-based survey, Landovitz et al. found that young MSM using Grindr reported high rates of sexual partnering and unprotected anal intercourse [1]. These results show that GSN apps have revolutionized social communication and partner-seeking among MSM, and therefore a different perspective based on GSN applications for understanding the MSM populations is required. However, empirical data in most studies is collected by survey-based methods with limited sample size, only few studies have focused on the social networking data generated on MSM-targeted GSN applications [21].

- *Mengsi Cai is with the College of Systems Engineering, National University of Defense Technology, Changsha, 410073, China. E-mail: caimengsi@nudt.edu.cn.*
- *Ge Huang is with the College of Economy and Management, Changsha University, Changsha, 410022, China. E-mail:gehuang_nudt@hotmail.com.*
- *Mirjam E. Kretzschmar is with Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, CX 3584, The Netherlands. E-mail: m.e.e.kretzschmar@umcutrecht.nl.*
- *Xiaohong Chen is with the School of Business, Central South University, Changsha, 410083, China, and with the Institute of Big Data and Internet Innovations, Hunan University of Technology and Business, Changsha, 410205, China. E-mail: cxhcsu@126.com.*
- *Xin Lu is with the College of Systems Engineering, National University of Defense Technology, Changsha, 410073, China. E-mail: xin.lu@flowminder.org.*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2021.3085984, IEEE Transactions on Network Science and Engineering

To fill this gap of knowledge, in this study, we collected a large-scale social networking dataset from the world's largest MSM-targeted GSN application named Blued, which includes 14,728,682 active users and 850,511,591 following relationships between users, covering the period from November 2012 to August 2018. Based on the dataset, we analyzed user profiles, such as age, location, and sexual role, to provide an overview of the MSM population at the global level. The "follow" relationship in Blued is primarily built on social ties, for the purposes of keeping in touch with others. Each user can freely follow others without others' agreement. Based on these following relationships, an MSM social network is constructed. In addition, we examined social network structures, such as degree distribution, reciprocity, degree assortativity, homophily, and community, to explore the online interaction patterns of the MSM population.

TABLE 1: Popular MSM dating applications in the world.

App	Founding year	Number of registered users
Blued	2012	40+ million (Dec 2018)
Hornet	2011	25+ million (Dec 2019)
Jack'd	2010	5 million (Dec 2019)
Scruff	2010	15+ million (Dec 2019)
Growlr	2010	10+ million (Dec 2019)
Grindr	2009	27 million (May 2017)
Romeo	2001	2 million (Dec 2019)

2 Materials and Methods

2.1 Data Collection

Blued, with more than 40 million registered users, is the world's largest application for MSM aged 17 and older. As a social networking application, Blued allows users to find partners, express opinions, and obtain social support; detailed activities include following other users, sending messages, editing personal profile information, setting partner preferences, publishing posts, voting viewpoints, making live broadcasting, joining groups, playing games, getting knowledge about HIV/AIDS, and ordering HIV tests. Also, the global positioning system (GPS) is utilized for subscribers to identify nearby users to facilitate the process of seeking partners. With the active users' daily activities and interactions on Blued, a large amount of social networking data has been accumulated. Based on this data, a large-scale dataset concerning the users and the follow-relationship-based social network was collected for this study by crawlers with Scrapy, a fast crawling framework in Python.

The user data is related to the registered users on Blued, mainly including the demographic information, such as the user's name, age, height, weight, sexual role, location, blood type, and ethnicity. Each registered user is identified by a unique *uid* number, which starts from 1 and increases by 1 along with the growth of the number of registered users. As a result, a *uid* set U = 1, 2, 3, …, n was constructed with an elaborate crawler, to reach all the registered users and extract their personal information from profile pages.

The social network data refers to the following relationships between users on Blued. Similar to the users in general online social networking platforms like Twitter and Facebook, Blued users can send and receive posts which can include text, pictures, or videos. Also, Blued users can follow other users to see their posts conveniently, which is beneficial for information transmission and communication. We use these following relationships to construct the MSM social network. Detailly, if user *i* follows user *j*, then we call user *i* as a follower of user *j*, and create a link from *i* to *j*. Similar to the user data collection, a crawler was developed to automatically visit the follower-list pages of all the users, and extract their following-relationships.

2.2 Data Preprocessing

With crawlers, we collected 14,728,682 user profiles and 850,511,591 following-relationships between them on Blued from November 2012 to August 2018. To construct a valuable dataset for further analysis, we removed the users who have withdrawn from Blued, because their personal information was reset to be NULL. Also, we removed the users with their age self-reported to be over 75, considering that older people usually have limited access to mobile applications. Finally, we obtained 838,910,078 following relationships among 14,668,867 active users. An active user is a person who accesses the Blued app and hasn't withdrawn this app during our observation period. All the user data and the following relationship data are saved in local MongoDB databases.

2.3 Network Analysis

A directed MSM social network was constructed from Blued data with users as nodes, and following relationships between users as edges. Only users who have at least one Blued friend are kept as nodes in the network. The network is presented as a directed graph, $G = (V, E)$, where V is the node set and E is the link set. $N = |V|$ represents the number of nodes in G, and $M = |E|$ denotes the number of links. The topological characteristics of the network are analyzed with several measures as follows.

In/out degree. The degree k_i in a network is the number of edges that are connected to the node *i*. The in-degree k_i^{in} is the number of edges directed to node *i*, and the out-degree k_i^{out} is the number of edges that node *i* directs to others. The average degree $< k >= (\sum k_i) / N$ indicates the average number of edges per node in the network.

Reciprocity. Link reciprocity describes the tendency of node pairs to form mutual connections between each other, which also reveals possible mechanisms of social network topology [22]. In this study, reciprocity R_i of node *i* is measured by the ratio of the number of neighbors with reciprocated links $L_i^{\leftrightarrow}$ (i.e. there is an edge in both directions between two nodes) to the total number of neighbors L_i (i.e. the number of unique users *i* is following or being followed): $R_i = L_i^{\leftrightarrow} / L_i$. Consequently, R_i reflects the bidirectional relationships between nodes. $R_i = 0$ means that all links of node *i* are directed, and $R_i = 1$ implies that all links are reciprocal. The average of all R_i is used to measure the global reciprocity of the network: $\overline{R} = (\sum_{i=1}^{N} R_i) / N$, in which N is the number of nodes in the network.

Degree assortativity. Degree assortativity is a preference

for a network's nodes to attach to others that are similar in degree [23]. Assortativity (or assortative mixing) indicates the tendency of network nodes to be connected to others with similar degree values; while disassortativity (or disassortative mixing) measures the tendency for nodes to connect to nodes with a higher or lower degree. The assortativity of network degree is measured by:

$$r = \frac{< k_{to} k_{from} > - < k_{to} > < k_{from} >}{\sqrt{k_{to}^2 - < k_{to} >^2} \sqrt{k_{from}^2 - < k_{from} >^2}} \quad (1)$$

in which k_{to} denotes the degree of the node that the edge points to, and k_{from} denotes the degree of the node that the edge starts from. This degree correlation function, which takes values in the range $-1 \le r \le 1$, is zero for no correlation, positive for assortative mixing, and negative for disassortative mixing. For the directed network in this paper, there are four combinations of degree assortativity, i.e. in-in degree assortativity r_{in-in}, in-out degree assortativity r_{in-out}, out-in degree assortativity r_{out-in}, and out-out degree assortativity $r_{out-out}$.

Homophily. Homophily is the probability that participants connect with friends who are similar to themselves regarding characteristics other than degree rather than connecting randomly [16]. Previous research has observed that human social networks display homophily in many sociodemographic, behavioral, and intrapersonal characteristics [24],[25], such as age, race, ethnicity, sex, religion, location, education, behavior patterns, and personality. As defined in accordance with Heckathorn [26], H = +1 means that all links are formed within the group, i.e. perfect homophily; H = 0 indicates that links are formed without regard to group members' attribute, i.e. no homophily; and H = -1 implies that all links are formed outside the group, i.e. heterophily. In summary, homophily's absolute value $| H |$ is the probability that homophily governs link formation.

Community. For many online activities, it has been shown that users tend to interact with others who are similar to themselves, forming distinct network communities, and stimulating studies on influence-based contagion or homophily-driven diffusion [27]. In this study, the community structure of the MSM social network is detected by Infomap [28], a highly efficient algorithm for detecting non-overlapping communities in directed weighted networks. Also, the cross-community links are analyzed to explore the organizational model and behavioral characteristics of the online MSM population.

3 Results

3.1 User Demographic

Demographics of 14,668,867 active Blued users aged 17 to 75 are presented in Fig. 1. The online MSM users are quite young, with a median (mean) age of 25 years (26.6 years). While 75.88% of users are in the younger age range between 17 and 29 years, only a small percentage (1.46%) of users are older than 50 years. Young MSM (age 17 to 29) and older MSM (age 50 and older) follow averagely 68.5 and 87.6 users, respectively, indicating a strong demand of older MSM for socializing and seeking partners. Regarding sexual roles, 22.27% of users identify as "versatile", while 18.76% and 14.31% identify as "top" and "bottom", respectively.

More than half of the users are from China (64.89%), with the rest registered from the other 216 different countries and regions (35.11%). A slightly smaller proportion (30%) of users outside China was reported by Blued in 2018, indicating the platform has since expanded its global coverage. In China, provinces with higher incidence of HIV such as Sichuan and Guangdong [29], are found to also have a high proportion of Blued users. The top 5 countries outside China with the highest numbers of Blued users were Thailand (775,230, 5.28%), the Philippines (607,102, 4.14%), Viet Nam (497,795, 3.39%), Indonesia (412,848, 2.81%), and Brazil (297,553, 2.03%). A majority of users are from the Asia-Pacific region, where the sexual minority population is often a marginalized and vulnerable group of society. In general, the data volume collected from GSN app analyzed in this paper is far greater than data sets collected using traditional survey methods in studies on MSM populations [30].

We use Pearson's chi-square goodness of fit test to evaluate if the distributions of users' age and sextual role in different countries are consistent with the age and sextual role distributions of all Blued active users. Results show that there is no significant difference among user age distributions across countries, and the distribution of sexual roles for users in 179 countries is consistent with the distribution of sexual roles for all Blued active users (see *Supplementary Materials*, Table S1, Fig. S1, and Fig. S2). This result to some extent indicates demographic similarities of MSM populations in different regions.

3.2 Degree Distribution

In this study, an MSM social network with 11,408,872 nodes and 838,910,078 edges was constructed based on the following relationships between Blued users. The average degree of the MSM social network is 73.5, which is much lower than that of social networks for general populations, such as Facebook (undirected) [31] and Twitter (directed) [32], but higher than of some social networks for heterosexual romantic relationships, such as the nioki.com network ($< k >= 8.07$) and the pussokram.com network ($< k >= 5.97$) [33]. A power-law distribution for degree is found in the MSM social network, as $p(k) \propto k^{-\alpha}$ where k denotes the in-/out- degree and α is the power-law exponent [34] (Fig. 2A and B). These results are similar to those reported in many other large human social networks as well as romantic social networks [35].

What is interesting about the nodal degree in the network is that the upper bound of out-degree ($k_{\max}^{out} = 127,448$) is relatively lower than the upper bound of in-degree ($k_{\max}^{in} = 206,095$), as also observed in the Twitter network [32], which indicates the limited capacity of information gathering. In addition, the in-degree and out-degree of individual users are extremely unbalanced, as shown in Fig. 2C and D. For example, users who follow thousands of others are found to be followed only by a few people, and users with thousands of followers only follow a few users. This might be explained to some extent by preferential

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2021.3085984, IEEE Transactions on Network Science and Engineering

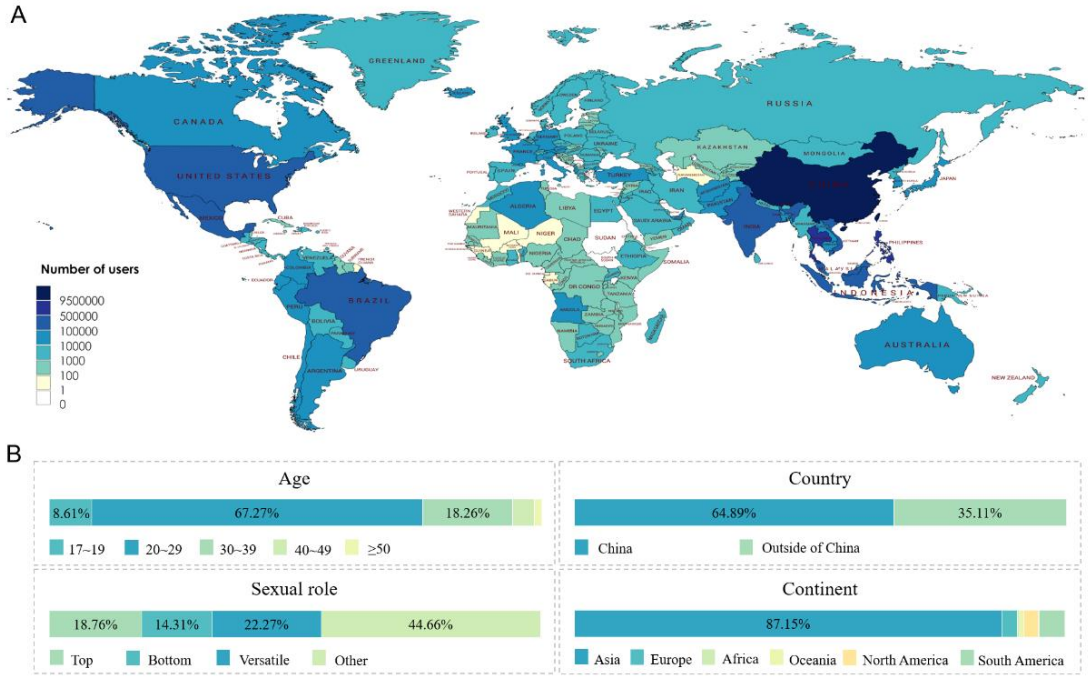


Fig. 1: Descriptions of Blues users. (A) The number of Blued users in each country. (B) User demographics statistics results of 14,668,867 active Blued users aged 17 to 75.

attachment, that users with high in-degrees are more visible and are therefore more likely to receive new followers. Such rich-club phenomena of degree distribution might also be influenced by the users' location, since we found that 69.05% of users with large in-degree (over 10k) and 84.96% of users with large out-degree (over 10k) are both located in China. In addition, as it has been reported that individuals are capable of maintaining only about 150 stable social relationships at a time in online and offline networks [36], the presence of users with thousands of following relationships is indicative of "non-social" behaviors.

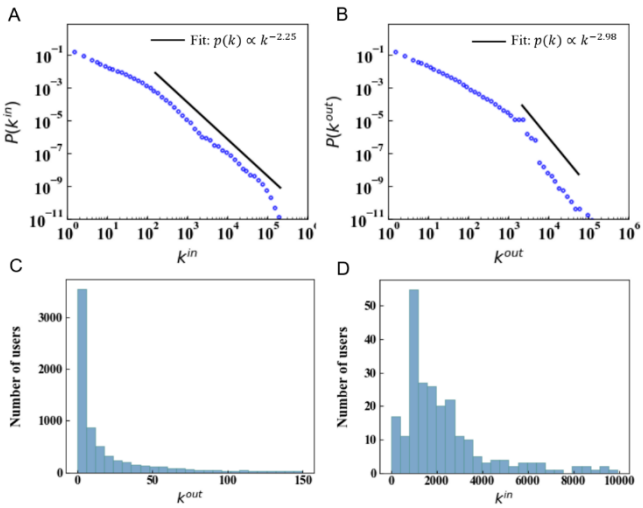


Fig. 2: Degree distributions. (A) In-degree distribution and (B) out-degree distribution of the MSM social network. The solid line represents power law fitting for the data. (C) Out-degree distribution of users with in-degree larger than 10,000. There are 7,674 users show an in-degree higher than 10,000, and 88.4% of them follow less than 150 users. (D) In-degree distribution of 236 users with out-degree larger than 10,000 and in-degree less than 10,000.

3.3 Reciprocity

Using the definition of reciprocity (see *Materials and Methods* for details), we measure the reciprocity R of all nodes in the MSM social network. Surprisingly, the average reciprocity of the network is only 0.047, which is much lower than that of heterosexual romantic social networks such as nioki.com (0.69) and pussokram.com (0.51) [33]. While 42% of edges in the Twitter follow graph are reciprocated [32], in the MSM social network, only 7.85% of edges are reciprocated, and the reciprocity for 59.50% of nodes is zero. One reason to expect a low reciprocity in the MSM social network is that a user can easily establish following relationships with strangers with the help of location-based services provided by Blued, and correspondingly he thus feels less social pressure to respond to a communicative stranger he is followed by. Another reason might be the high turnover rate of friends, which has been found in a large online MSM dating community in China [37]. That is, users might frequently change (establish or delete) their following relationship during a short time.

Fig. 3 demonstrates the average reciprocity with respect to nodes with increasing numbers of neighbors, increasing age, and by country. As shown in Fig. 3A, when the number of neighbors of a node exceeds 1000, the reciprocity becomes smaller and shows a slight declining tendency. This suggests that people can only maintain a small percentage of reciprocal relationships when they have a large number of friends, as also observed in mobile communication networks [38]. Interestingly, as revealed by Fig. 3B, users aged 38 to 55 have a higher reciprocity than others, suggesting their strong demand and great efforts to making friends.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2021.3085984, IEEE Transactions on Network Science and Engineering

Furthermore, the average reciprocity for nodes in most countries is around 0.04 (Fig. 3C). Users with different sexual roles also show similar average reciprocity, the average reciprocities for the tops, the bottoms, and the versatiles are 0.039, 0.041, and 0.045, respectively.

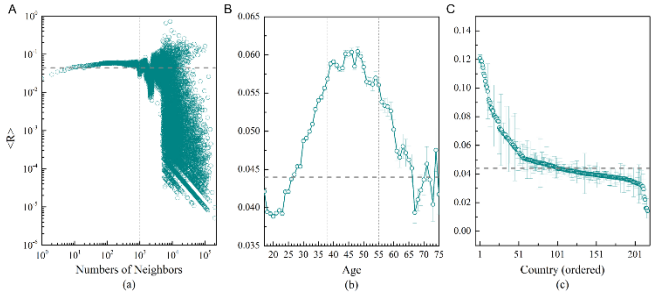


Fig. 3: Distribution of average reciprocity *R* for nodes with different (A) numbers of neighbors, (B) ages, and (C) located countries ordered by the average reciprocity <*R*>. The gray horizontal dash line indicates the average reciprocity of the network (0.044).

3.4 Degree Assortativity

As presented in Table 2, we compared the degree assortativity of the directed MSM social network with heterosexual romantic social networks (i.e. the pussokram.com network), general social networks (i.e. Twitter and Facebook), and collaboration networks [31],[33],[39]. In addition, considering that the very low level of reciprocity might be caused by that a link does not imply an actual interaction, we further studied the undirected MSM social network which is constructed by only the reciprocal following relationships.

A positive r_{in-out} is found in the directed MSM social network, which is similar to the Twitter network and indicates that popular nodes (with larger in-degree) connect with nodes of larger out-degree, probably driven by preferential attachment. For the undirected networks, it is interesting that romantic social networks like Blued and pussokram.com are disassortative, while a positive degree-degree correlation between 0.1 and 0.4 is typically found in the Facebook network and in collaboration networks [23]. This implies that users in romantic social networks with large popularity or activity tend to connect with low-degree users. These results can be explained by the findings on exchange networks that, disassortative mixing is optimal when individuals are substitutable in forming particular relationships [40], as for example in friendship and dating networks. In contrast, assortative mixing is observed in collaboration networks because potential alternative collaborators have been already screened off [33]. The disassortative mixing in Blued could also be explained by the quick change of friends or the preference of making new friends, that is, a negative degree assortativity might be produced when highly active users frequently contact with strangers who are probably new or less connected Blued users with lower degrees.

3.5 Homophily

We divided the Blued users into several groups by age, location, and sexual role, and measured the homophily H_{g_i} for different groups g_i in both directed and undirected MSM social networks with the definition in [26]. As shown in Table 3, in the directed MSM social network, the homophily for young MSM aged 17 to 29 years is positive (0.33), indicating that young MSM tend to establish friendships with each other. An interesting age-preference friendship in the network is revealed in Fig. 4A, that is, 76.8% of the edges target young MSM, and users of all ages are most likely to follow young MSM users. In addition, users show similar following patterns for different age groups, which is very different from Facebook, where users mostly follow people with a similar age [30]. However, the homophily for young MSM in the undirected MSM social network is negative (-0.53), which implies that young MSMs tend to establish a reciprocal relationship with people who are older than themselves.

In terms of sexual roles, the homophily for the tops in the directed network is close to zero, indicating that links are formed by this group's members with no preference for others' sexual roles. Users who identify as "top" are the most popular group for all groups including the tops, the bottoms and the versatiles, since $p(top|top)$, $p(top|bottom)$ and $p(top|versatile)$ are fairly high, as shown in Fig. 4B. It is straightforward that the negative homophily among the bottoms (-0.29 and -0.75) is due to the fact that a bottom tends to contact with others who are different from themselves in sexual role, such as a top or a versatile. It also reveals the bottom's disadvantage of being selected as friends by others. And it is interesting that the homophily for all the sexual role groups are negative, which indicates that reciprocal relationships are formed between users with different sexual role.

Regarding the users' locations, we found that in the directed MSM social network, 74.11% of edges are established between users in the same countries, and 60.33% of edges are formed between China-located users. The homophily for countries with more than 10,000 users ranges from 0.0016 to 0.82, with a median (average) of 0.19 (0.27). A high positive homophily of 0.53 in China-located users indicates that in addition to interacting with other users randomly, there is 53% surplus of tendency for China-located users to build interactions with people who are also from China. And as observed in Fig. 4C, users tend to establish friendships with people in the same geographic location (the diagonal of the figure), and the China-located users have a high possibility to be a neighbor of others for their large population (the first column of the figure). Similar results can be found in the undirected MSM social network. In the same way as people in other social networks [24], online MSM users in homosexual romantic social networks also consider the similar geographic location as an important condition when making friends and seeking partners, in order to make arrangements for off-line dating in reality.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2021.3085984, IEEE Transactions on Network Science and Engineering

TABLE 2: Degree assortativity for different social networks.

Type	Social networks	$r_{in\text{-}in}$	$r_{in\text{-}out}$	$r_{out\text{-}in}$	$r_{out\text{-}out}$	r
Directed	Blued	-0.064	0.016	-0.001	-0.002	/
	Pussokram. com [33]	-0.063	-0.046	-0.071	-0.050	/
	Twitter [32]	-0.296	0.241	-0.118	0.272	/
Undirected	Blued (only reciprocal links)	/	/	/	/	-0.0005
	Pussokram. com [33]	/	/	/	/	-0.048
	Facebook [31]	/	/	/	/	0.226
	Physics coauthorship [39]	/	/	/	/	0.363
	Biology coauthorship [39]	/	/	/	/	0.127
	Mathematics coauthorship [39]	/	/	/	/	0.120
	Film actor collaborations [39]	/	/	/	/	0.208
	Company directors [39]	/	/	/	/	0.276

TABLE 3: Homophily H for different groups g_i in the directed and undirected MSM social network.

Network type	Group g_i	Age		Country		Sexual role			
		≤ 29	> 29	China	outside China	top	bottom	versatile	Other
Directed	H	0.33	0.15	0.53	0.70	0.04	-0.29	-0.17	0.19
Undirected	H	-0.53	0.07	0.49	0.53	-0.50	-0.75	-0.33	-0.08

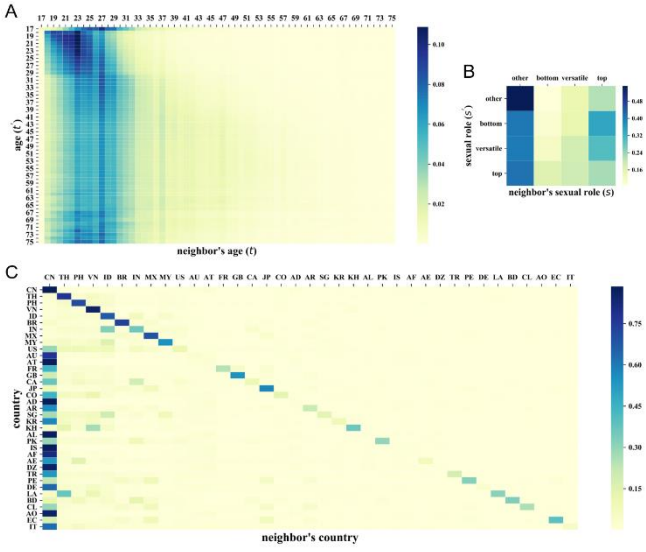


Fig. 4: Friendship patterns between different ages, sexual roles and locations. (A) The distribution $p(t'|t)$ of age t' for the neighbors of users with age t. $p(t'|t)$ is the conditional probability of individuals with age t (t =17, 18, …, 75) selecting a random neighbor who has age t' (t' =17, 18, …, 75). The color represents the value of $p(t'|t)$ in each square. (B) The distribution $p(s'|s)$ of sexual role s' for the neighbors of users with sexual role s. $p(s'|s)$ is the conditional probability that a random neighbor of individuals with the sexual role of s (s = other, bottom, versatile, top) plays the sexual role of s' (s' = other, bottom, versatile, top). The color represents the value of $p(s'|s)$ in each square. (C) The distribution $p(c'|c)$ of country c' for the neighbors of users from country c. $p(c'|c)$ is the conditional probability that a random neighbor of individuals from country c. The color represents the value of $p(c'|c)$ in each square. Countries with the number of active users over 10,000 are shown in the figure. The complete list of countries presented in the heatmap and ordered by the number of users in each country is shown in *Supplementary Materials*, Table S2.

3.6 Community

Using the Informap algorithm [28], a total of 26,426 communities are detected from the MSM social network as shown in Fig. 5. The community size ranges from 1 to 8,229,292, with a majority of communities containing only one node (62.5%), and the mean size is 432. While the numbers of nodes in 98.23% of the communities are less than 10, the largest community contains 72.13% of the nodes in the network. The distributions of users' ages, locations, and sexual roles are similar among communities with more than 1,000 nodes, indicating that different users are mixed in communities and the communities are not simply divided by users' demographic attributes.

There are 168,628,787 (20.1%) edges in the network that are cross-community links connecting individuals in different communities, which suggests an active interaction between different network communities. Based on the formation of cross-community links, we extract two kinds of nodes: inner nodes that are not associated with any cross-community links ($\odot$), and boundary nodes that are associated with both cross-community out-links and in-links ($\oplus \to \odot \to \oplus$). The proportion of boundary nodes in the network is 36.68%, that is, one-third of the nodes interact closely with nodes in different communities. Also, the degree of boundary nodes ($k^{in}=129.7$, $k^{out}=141.7$) is greater than the average degree of the network (73.5) and far greater than the degree of inner nodes ($k^{in}=5.3$, $k^{out}=2.2$), indicating that the boundary nodes are quite active and popular in the network and they play an important role in maintaining global connectivity between communities. As

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TNSE.2021.3085984, IEEE Transactions on Network Science and Engineering

shown in Fig. 6, it is obvious that the two largest communities have larger maximum in-/out- degrees and more inner/ boundary nodes than small communities. With the decrease of the rank of community size, there are slight declines in the maximum in-/out- degree and the number of inner/ boundary nodes. However, no correlation is found between the community size and average degrees.

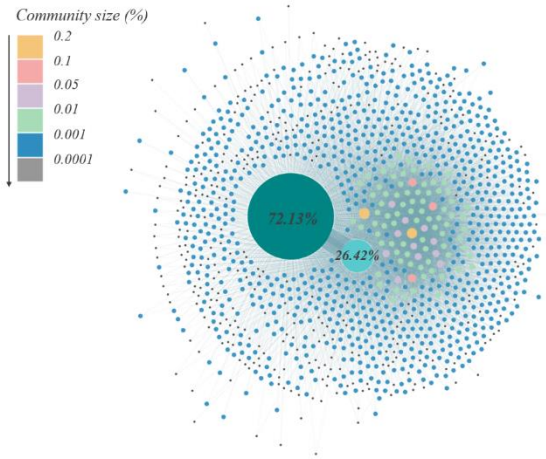


Fig. 5: Distribution of the communities' size and the interactions between communities in the MSM social network. The size of a community is the number of nodes in the community. Each node represents a community. The size of the node is proportional to the number of users in each community. Only communities with at least 5 users are displayed in the figure.

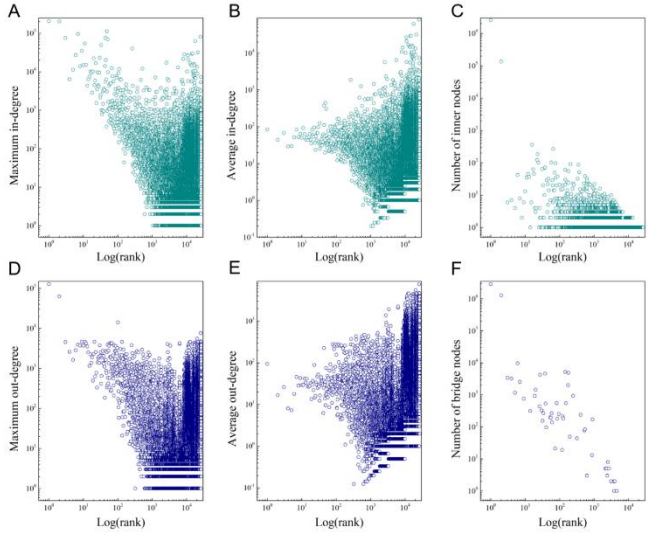


Fig. 6: Distributions of (A) maximum in-degree, (B) average in-degree, (C) number of inner nodes, (D) maximum out-degree, (E) average out-degree, and (F) the number of boundary nodes in all communities in log-log scale. The circle in the figures indicates the community. The horizontal axis represents the rank of communities, which is sorted by community size in descending order.

4 Discussion

With 14.7 million users and 850 million edges collected from the world's largest MSM-targeted GSN app named Blued, the dataset used for our analysis on the user demographics and social network structure is, to our knowledge, the largest sample of MSM at the global level. Young MSM aged 17 to 29 years who are reported having the highest HIV incidence [1],[41], are active online and constitute 75.88% of the Blued users. Older MSM aged over 50 years are found to have a strong demand for socializing and partner seeking, while little research up to now has focused on this age group due to ideas of asexuality associated with advanced age and the difficulty in recruiting these subjects for research studies [41][42].

In the MSM social network, most users have less than 300 following relationships, while a few have following relationships with thousands of other users. Though an MSM user has an average of 73.5 friends, they seem to fail to establish reciprocal friendships with others. Furthermore, users in Blued are found to usually interact with strangers online[17]. An interesting finding is that the network shows extremely low reciprocity and negative degree assortativity. This is opposite to what is found in general social networks and collaboration networks, but could be explained by a high turnover rate of close friends on Blued for dating and one-night stands in reality. The differences reported between romantic social networks (i.e. Blued for homosexual relationships and pussokram.com for heterosexual relationships) and general social networks (i.e. Facebook and Twitter) suggest that research efforts should be made to uncover the underlying mechanism in the formation of romantic social relationships from different perspectives such as the extremely low reciprocity or the high turnover rate of close friends.

Another interesting finding is, that the network shows a strong homophily on geographic location and a clear preference of age selection. That is, close geographic location becomes an important condition considered by MSM users when socializing, for the purpose of off-line dating, and users in all age groups are most likely to seek contact with young MSM users. When dividing the network into several communities, an extremely large community containing 72.13% of the users in the network is found, but users in different communities present similar distributions in age, location, and sexual role.

Using the massive data set generated online, we have obtained data from the MSM population of an unprecedented data volume. In comparison with the data collected by survey-based methods, the use of social networking big data of MSM on GSN apps has many advantages. First, GSN apps contain millions of MSM users from all around the world, which is difficult or even impossible to be collected via survey. Second, the relationships between MSM users recorded by GSN apps are more objective and reliable than those collected by survey-based methods. For example, survey results are typically subjected to memory bias or reporting errors due to sensitivity and privacy concern of respondents. Third, large scale multilingual text has been generated in social interactions on GSN apps, which can be used to infer the emotional status and topics they discuss online[17]. Finally, we revealed many social behavior characteristics in this paper which have never been suggested by survey-based studies, for example, MSM users in their twenties are excessively connected by

users from all age groups and "top" MSM users are more popular than "bottoms" and "versatiles" in social networking.

However, as this study used all the active users on Blued without identifying whether they are MSM, the representativeness of the sample remains to be discussed. Another limitation with this study is the possible lack of reliability of self-reported data, which remains a common limitation of Internet data. Future research will focus on the representativeness of online data and the exploration of temporal and spatial patterns of users' online and offline social behaviors.

5 Conclusion

Men who have sex with men, a population who is active in online social networking applications but hard to access for research, form a highly vulnerable community for HIV acquisition. However, previous survey-based studies are mostly limited in sample size and accessibility. In an effective manner, in this study we obtained a large-scale social networking dataset that contains 11,408,872 active users and 838,910,078 user-following relationships from the world's largest MSM geosocial networking application named Blued. We presented quantitative social networking analysis of the users and the social network based on their following relationships, showing how MSM social networks differ from other social networks, including heterosexual social networks and general social networks. In addition, we demonstrated a new way for studying the MSM population via online social networking applications, which is effective in reaching MSM communities. Insights from these analyses can help inform health interventions in MSM populations.

Acknowledgment

XL is supported by the National Natural Science Foundation of China (91846301, 71771213, 72025405), MC is supported by the National Natural Science Foundation of China (82041020, 71690233, 71790615). This study was also supported by the Hunan Science and Technology Plan Project (2019GK2131, 2020TP1013). The authors declare no conflicts of interest. Xin Lu is the corresponding author.

Data Access Statement

The datasets and codes used in this study are available online at: (link will be available upon publication).

References

[1] RJ Landovit, et al., "Epidemiology, sexual risk behavior, and HIV prevention practices of men who have sex with men using Grindr in Los Angeles, California," *J. Urban Health*, vol. 90, no. 4, pp. 729-739, 2013.

[2] R Magnani, K Sabin, T Saidel, and D Heckathorn, "Review of sampling hard-to-reach and hidden populations for HIV surveillance," *AIDS*, vol. 19, no. S2, pp. S67-S72, 2005.

[3] AM Bowen, K Horvath, and ML Williams, "A randomized control trial of internet-delivered HIV prevention targeting rural MSM," *Health Educ. Res.*, vol. 22, no. 1, pp. 120-127, 2007.

[4] L Bengtsson, X Lu, F Liljeros, HH Thanh, and A Thorson, "Strong propensity for HIV transmission among men who have sex with men in Vietnam: Behavioural data and sexual network modelling," *BMJ Open*, vol. 4, no. 1, pp. e003526, 2014.

[5] KE Tobin, M Cutchin, CA Latkin, and L Takahashi, "Social geographies of African American men who have sex with men (MSM): A qualitative exploration of the social, spatial and temporal context of HIV risk in Baltimore, Maryland," *Health Place*, vol. 22, no. C, pp. 1-6, 2013.

[6] K Madkins, et al., "Attrition and HIV risk behaviors: A comparison of young men who have sex with men recruited from online and offline venues for an online HIV prevention program," *Arch. Sex. Behav.*, vol. 47, no. 1, pp. 2135-2148, 2018.

[7] Y Guo, and DHL Goh, "'I Have AIDS': Content analysis of postings in HIV/AIDS support group on a Chinese microblog," *Comput. Human Behav.*, vol. 34, pp. 219-226, 2014.

[8] PK Mo and NS Coulson, "Exploring the communication of social support within virtual communities: A content analysis of messages posted to an online HIV/AIDS support group," *Cyberpsychol. Behav.*, vol. 11, no. 3, pp. 371-374, 2008.

[9] C Oldenburg, et al., "Structural stigma affects access to pre-and post-exposure prophylaxis and HIV risk among men who have sex with men (MSM) in the United States," *AIDS Res. Hum. Retroviruses*, vol. 30, no. S1, pp. A22-A23, 2014.

[10] SA Safren, AJ Blashill, and CM O'Cleirigh, "Promoting the sexual health of MSM in the context of comorbid mental health problems," *AIDS Behav.*, vol. 15, no. S1, pp. 30-34, 2011.

[11] M Marpsat and N Razafindratsima, "Survey methods for hard-to-reach populations: Introduction to the special issue," *Method. Innov.*, vol. 5, no. 2, pp. 3-16, 2010.

[12] J Larmarange, et al., "Men who have sex with men (MSM) and factors associated with not using a condom at last sexual intercourse with a man and with a woman in Senegal," *PLoS One*, vol. 5, no. 10, pp. e13189, 2010.

[13] X Lu, L Bengtsson, T Britton, M Camitz, BJ Kim, A Thorson, and F Liljeros, "The sensitivity of respondent-driven sampling," *J. Roy. Stat. Soc. Stat. Soc.*, vol. 175, no. 1, pp. 191-216, 2012.

[14] X Lu, J Malmros, F Liljeros, and T Britton, "Respondent-driven Sampling on Directed Networks," *Electron. J. Stat.*, vol. 7, pp. 292-322, 2013.

[15] D Juher, J Saldaña, R Kohn, K Bernstein, and C Scoglio. "Network-centric interventions to contain the syphilis epidemic in San Francisco," *Sci. Rep.*, vol. 7, no. 1, pp. 1-17, 2017.

[16] X Lu, "Linked ego networks: improving estimate reliability and validity with respondent-driven sampling," *Soc. Networks*, vol. 35, no. 4, pp. 669-685, 2013.

[17] G Huang, M Cai, and X Lu, "Inferring opinions and behavioral characteristics of gay men with large scale multilingual text from Blued," *Int. J. Environ. Res. Public Health*, vol. 16, no. 19, pp. 3597, 2019.

[18] HN Czarny and MR Broaddus, "Acceptability of HIV prevention information delivered through established geosocial networking mobile applications to men who have sex with men," *AIDS Behav.*, vol. 21, no. 11, pp. 3122-3128, 2017.

[19] CH Bien, et al., "Gay apps for seeking sex partners in china: implications for MSM sexual health," *AIDS Behav.*, vol. 19, no. 6, pp. 941-946, 2015.

[20] G Phillips, et al., "Use of geosocial networking (GSN) mobile phone applications to find men for sex by men who have sex with men (MSM) in Washington, DC," *AIDS Behav.*, vol. 18, no. 9 pp. 1630-1637, 2014.

[21] B Zhao, DZ Sui, and Z Li, "Visualizing the gay community in Beijing with location-based social media," *Environ. Plan. A*, vol. 49, no. 5, pp. 977-979, 2017.

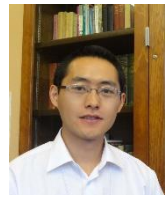
[22] D Garlaschelli and MI Loffredo, "Patterns of link reciprocity in directed networks," *Phys. Rev. Lett.*, vol. 93, no. 26 Pt 1, pp. 268701, 2004.
[23] ME Newman, "Mixing patterns in networks," *Phys. Rev. E*, vol. 67, no. 2, pp. 26126, 2003.
[24] N Noë, RM Whitaker, MJ Chorley, and TV Pollet, "Birds of a feather locate together? Foursquare checkins and personality homophily," *Comput. Human Behav.*, vol. 58, no. C, pp. 343-353, 2016.
[25] N Noë, RM Whitaker, and SM Allen, "Personality homophily and geographic distance in Facebook," *Cyberpsychol. Behav. Soc. Netw.*, vol. 21, no. 6, pp. 361-366, 2018.
[26] DD Heckathorn, "Respondent-driven sampling II: Deriving valid population estimates from chain-referral samples of hidden populations," *Soc. Probl.*, vol. 49, no. 1, pp. 11-34, 2002.
[27] C Liu and X Lu, "Analyzing hidden populations online: topic, emotion, and social network of HIV related users in the largest Chinese online community," *BMC Med. Inform. Decis. Mak.*, vol. 18, no. 1, pp. 1-10, 2018.
[28] M Rosvall and CT Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proc. Natl. Acad. Sci.*, vol. 105, no. 4, pp. 1118-1123, 2008.
[29] The data-center of china public health science (2019) http://www.phsciencedata.cn/Share/.
[30] C Maulsby, et al., "HIV among black men who have sex with men (MSM) in the united states: a review of the literature," *AIDS Behav.*, vol. 18, pp. 10-25, 2014.
[31] J Ugander, B Karrer, L Backstrom, and C Marlow, "The anatomy of the Facebook social graph," *arXiv preprint*, arXiv:1111.4503, 2011.
[32] SA Myers, A Sharma, P Gupta, and J Lin, "Information network or social network? The structure of the twitter follow graph," *The 23rd International Conference on World Wide Web*, 2014, pp. 493-498.
[33] P Holme, CR Edling, and F Liljeros, "Structure and time evolution of an internet dating community," *Soc. Networks*, vol. 26, no. 2, pp. 155-174, 2004.
[34] J Alstott, E Bullmore, and D Plenz, "Powerlaw: A python package for analysis of heavy-tailed distributions," *PLoS One*, vol. 9, no. 1, pp. e85777, 2014.
[35] F Liljeros, CR Edling, LAN Amaral, HE Stanley, and Y Åberg, "The web of human sexual contacts," *Nature*, vol. 411, no. 6840, pp. 907-908, 2001.
[36] RI Dunbar, "Do online social media cut through the constraints that limit the size of offline social networks?", *R. Soc. Open Sci.*, vol. 3, no. 1, pp. 150292, 2016.
[37] C Liu and X Lu, "Network evolution of a large online MSM dating community: 2005-2018," *Int. J. Environ. Res. Public Health*, vol. 16, no. 22, pp. 4322, 2019.
[38] Q Wang, J Gao, T Zhou, Z Hu, and H Tian, "Critical size of ego communication networks," *Europhys. Lett.*, vol. 114, no. 5, pp. 58004, 2016.
[39] ME Newman, "Assortative mixing in networks," *Phys. Rev. Lett.*, vol. 89, no. 20, pp. 208701, 2002.
[40] KS Cook, RM Emerson, MR Gillmore, and T Yamagishi, "The distribution of power in exchange networks: Theory and experimental results," *Am. J. Sociol.*, vol. 89, no. 2, pp. 275-305, 1983.
[41] J Prejean, et al., "Estimated HIV incidence in the United States, 2006-2009," *PLoS One*, vol. 6, no. 8, pp. e17502, 2011.
[42] C Daas, M Doppen, AJ Schmidt, and EO Coul, "Determinants of never having tested for HIV among MSM in the Netherlands," *BMJ Open*, vol. 6, no. 1, pp. e009480, 2016.

Mengsi Cai received the B.E. degree and M.E. degree in information science from Xiangtan University, Xiangtan, Hunan, China, in 2014 and 2017, respectively. She is currently pursuing the Ph.D. degree in management science and engineering with the National University of Defense Technology, Changsha, Hunan, China. She is a yearly visiting researcher at University Medical Center Utrecht in the Netherlands since December 2019. Her research interests include complex networks, data mining, and natural language processing.

Ge Huang received the Ph.D. degree in management science and engineering from the National University of Defense Technology, Changsha, Hunan, China, in 2019. She is currently an instructor of management in Changsha University, Changsha, Hunan, China. Her research interests include data mining, natural language processing, and complex networks.

Mirjam E. Kretzschmar studied Mathematics and Biology at Johannes Gutenberg University in Mainz and Eberhard Karls University in Tübingen (Germany). She obtained her doctorate with the highest possible distinction for her research into models for parasitic infections. From 1987 to 1992, she worked at research institutes in Germany, the Netherlands, the United States and the United Kingdom. She is currently a professor of Dynamics of Infectious Diseases at the Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, the Netherlands. And she is Chief Science Officer (CSO) of Mathematical Disease Modelling at RIVM (National Institute for Public Health and the Environment).

Xiaohong Chen received the Ph.D. degree in Engineering from Tokyo Institute of Technology, Tokyo, Japan, in 1999. She is currently an academician of the Chinese Academy of Engineering (CAE), president of Hunan University of Technology and Commerce, honorary dean of Business School of Central South University. She currently serves as a committee member of the National Natural Science Foundation of China and a committee member of Discipline Assessment Group of Degree Committee of State Council. Her research interests include decision theory and decision support system, big data analysis, SME financing, resource conserving & environment friendly society and ecological civilization.



Xin Lu obtained his Ph.D. degree in Medical Science from Department of Public Health Sciences at Karolinska Institutet in 2013. He is currently a professor at the National University of Defense Technology in China. He has been granted the National Science Fund for Distinguished Young Scholars. His research interests include big data analytics, complex networks, human behaviour and emergency management.