Predicting offline behaviors from online features — an ego-centric dynamical network approach

Lianghao Dai
Department of Sociology
Tsinghua University
Beijing, China
anaedlh@gmail.com

Jar-der Luo
Department of Sociology
Tsinghua University
Beijing, China
jdluo@mail.tsinghua.edu.cn

Xiaoming Fu
Inst. of Computer Science
University of Goettingen
Goettingen, Germany
fu@cs.uni-goettingen.de

Zhichao Li
Department of Sociology
Tsinghua University
Beijing, China
yizhiqingcao@163.com

ABSTRACT

Investigating online social behaviors may help us to better understand and predict offline high risk behaviors in gay communities. But how can offline behaviors be predicted from online social networks? This article selects data from 26 online social network groups from QQ (a Chinese based messaging software) administered by gay communities of "W" city of Hubei Province, China. Based on online data mining, social network analysis, and offline semi-structural interviews, we argue that the ego-centric dynamical network analysis—an approach which combines partial network dynamics, individual features, and structure position together—can be used to derive the probabilistic features for predicting offline high risk behaviors (HRB). An example of HRB is "one night stands" (gays for one night: 419) for gay homosexuals.

Keywords

Online-offline social networks, ego-centric dynamical network analysis, high risk behavior, homosexual, QQ

1. INTRODUCTION

Prediction of offline behavior from online data is a young topic. Although a rich amount of work has been done on simulated online network structural evolutions, network ethnography, and all other kinds of statistic analysis between online and offline networks, there are few efforts addressing the issue of predicting offline behaviors using online data. In this paper we classify the related research into two categories:

First, theoretical models are used to describe properties of social networks. Much work has been focused on testing power law degree distributions [1] [2], and small world properties [3] in various networks. Besides, Kumar et al. argue that there are three groups of users within Flickr and Yahoo!'s 360 online social networks: singletons, the giant component, and the rest [4]. They also find that each group has their own evolution structure. However, the above-mentioned studies are mostly based on simulations which assume certain interaction patterns to test specific models or properties of networks. As a result, they can at most approximately understand how the network evolution process works online, but cannot more accurately depict the corresponding offline relationships and behaviors that occur accordingly.

Second, there are also empirical studies (such as cyber-anthropology [5]) that focus on identification building, strategies, and relationships in interacting processes between online and offline networks. For example, Khaled and et al. find that there are several other roles for users in online communities such as "lurkers" and "active participators" [6]. Ploderer, et al. found that there are three kinds of interacting motivations (using online social network as a tool, a theatre, and a community) among users [7]. Also, Subrahmanyam et al. studied the overlaps between participates online and offline networks and found that the emerging adults use different kind of online contexts to strengthen offline ties. These empirical studies are strictly based on interviews and participation observations. As such they do not take the relationship with online networks into account, and the time dimension for quantity description and prediction is fairly limited.

We argue that neither mathematical models on online social networks nor online-offline investigations are enough to realize the goal of effective prediction of offline behaviors. To achieve this goal, we develop a combined approach which take advantages of both, and present our use case study of ego-centric dynamical network analysis on gay homosexual groups in the "W" city, China. We select this target group for the following reasons: 1) It is a group with a clear goal—searching gay friends — with rich accessible online discussion records/data. 2) In our preliminary study of gay groups within one of the most popular online social network sites, their frequent offline group behaviors can be clearly observed or described through online discussions. 3) There are multiple relationships among the members of the groups, which indicate their networks are far more complex than the groups in only one kind of relationship. We conclude this paper by giving several possible hypotheses for predicting high risk behaviors (HRB) of the gay homosexuals in the "W" city.

2. DATA SET AND METHODOLOGY

2.1 Data Set

We select, by Convenient Sampling [8], 26 QQ groups within 1893 users who had online discussion activity[1], from a local NGO

ACM HotSocial'12, August 12, 2012, Beijing, China.

[1] The exact number of whole QQ gay group users is unaccountable because of high probability of online population mobility. However, there are almost 5000-7000 ID numbers that have ever been in these groups by roughly estimation.

(None Governance Organization) working on HIV protection and Gay community cultivation. We used web chat logs from May 2011 to April 2012. For this investigation, we chose to use the QQ software because it is the most widely used media tool for local Chinese homosexuals. However, a whole and complete number is beyond calculation because of the elusiveness of this topic within China. The NGO was able to observe at least 200 groups.

Based on our time, we conducted 8 semi-structural interviews [8]. They are selected from and based on multiple online structural positions in their chatting networks (As their positions are not steady, we do not list them here as interdiction. However, all four positions—Key nodes, active groups, normal nodes, and bubbles & lurkers—are included in our interviewees), jobs, ages, duration of years in the community, self identity (play a passive (feminized) or positive (masculine) role in intimate relationship) and dating environment.

Interviewee 1: Int*erior designer; 23 (years old); 4th year in community; passive role.*

Interviewee 2: NGO stuff; 20; 7th year; positive role

Interviewee 3: College student; 20; 2nd year; passive role

Interviewee 4: College student; 25; 6th year; positive role

Interviewee 5: Bus driver; 27; 4th year; positive role

Interviewee 6: NGO stuff; >30; >7th year; positive role

Interviewee 7: NGO stuff; 30; >7th year; positive role

Interviewee 8: College student; 20; 2nd year; passive role

Figure 2.1 A QQ Group

2.2 Related work on homosexuals

There are a few relevant literatures about online homosexuals' behaviors and their trigging of offline group behaviors. Jones et al. [9] studied the discourse of sex and identities of gays in Cyberspace, and found that in online communities of text-based sex, users may transgress the bounds of bodilessness by sending photos, phone calls, and ultimately, face-to-face, though not all the face-to-face dating led a positive result to their relationships. On the other hand Campbell [10] showed that offline social positions of gay individuals also shape online social dynamics. Besides, Goldman [11], from the aspect of individual, argued that for online daters, both men and women (who may not be homosexuals) routinely strive for the "prettiest, best partner out there." Also, once those two virtually smitten people finally meet and get past any initial—and one hopes, not insurmountable—disappointment, physical attractiveness is the least important factor when it comes to predicting the likelihood of a second date. All aforementioned literature are on dating, one-to-one relationships rather than group behaviors, yet the individual factors of social positions (which means interactions of online and offline networks), physical attractiveness, and desires for best partners are indeed quite remarkable among gay groups. However, none of the authors examined the behaviors from the viewpoint of social networks, which plays an important role.

2.3 Online Data Mining

Users in QQ groups have their conversations by broadcasting, which means everyone in the group can see their logs. The log structure for each participator is shown below:

User name (User ID) Time (hh:mm:ss)

[Chatting texts]

However, for each log, only a few individuals react to chat thus build ties between them. So we define the effective chatting relationships by a section of conversations in a X-second window. We find that participators place their attentions on brand new messages after an average of 15 seconds. Here we set X=15, so that one builds chatting relationships those who post in the 15 seconds both before and after one's own log.

In this way we translate QQ log context into chatting relationship arrays like {user1, user2, chatting frequency} then build them into a users-users chatting frequency matrix in every half month. We set this time duration as our observation point because features do not display steady frequency in a shorter duration periods. Based on the matrices, we calculate the degree distributions and features of the small world phenomenon. We also visualize the matrices in graphs with the help of Ucinet.

2.4 Offline Interviews and Participating Observations

For better understanding of offline behaviors of homosexuals, it is hard to use questionnaires to measure offline behaviors because we cannot expect interviewees will give reliable answers to a strange interviewer. However, semi-structured interviews [8] with friendly and relaxed face-to-face interactions are conducted with people within these groups. Internet using habits, personal stories of dating from online to offline, and possible factors of mate selection are asked. Besides, more than 1 month long participating observation is also held.

3. RESULT

3.1 Structure Positions

These 26 groups show international topology structure. The graphs also tell us there are four structural roles in this online community. Each of the four roles have their own behavior patterns, situations in mate selection, and probability of HRB. Compared with this classification, Khaled and his colleagues' categories are quite simplistic.

1) Key nodes

Key nodes include bridges and the nodes with max degrees. They are important not only because they play the role of structure holes [12] and opinion leaders, but also the phenomenon that members of this kind of key nodes change every half of month in this group (shown in Figure 3.1). It is obviously an astonishing finding because key nodes are very important in aggregating members in the community.

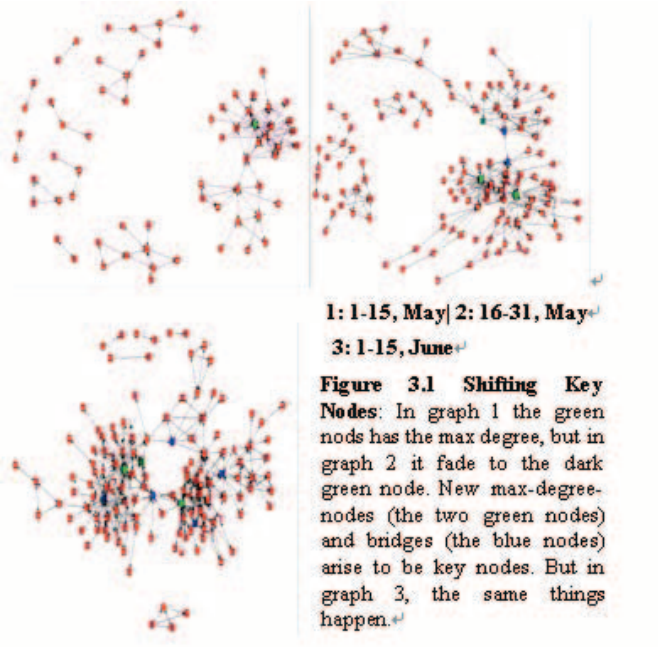


Figure 3.1 Shifting Key Nodes: In graph 1 the green nods has the max degree, but in graph 2 it fade to the dark green node. New max-degree-nodes (the two green nodes) and bridges (the blue nodes) arise to be key nodes. But in graph 3, the same things happen.

2) Participators in active small groups

Fig 3.2 shows that there are several structured steady active groups during the long period of time. Though particular individuals change frequently, these groups play a role of maintaining both online and offline relationships.

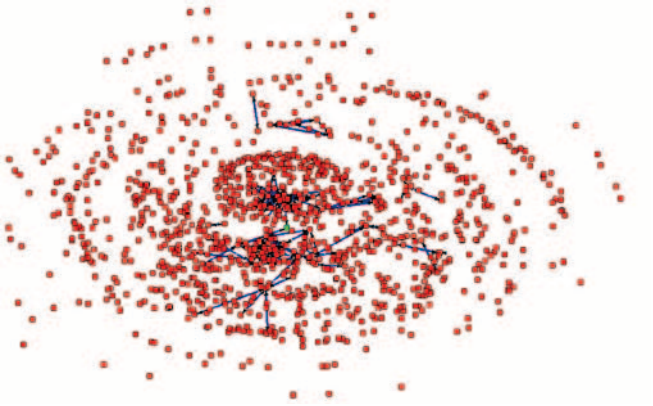


Figure 3.2 Leaving Edges with f>54, Small Active Groups can be Seen.

3) Lurkers and Bubbles

Lurkers are singletons in Khaled and his colleagues' categories. They are 0-fequency nodes. On the contrary, bubbles are nodes with a few frequencies (1-3 as usual). As they have very different behavior patterns and huge population compared with others, we pay an attention to these roles.

4） Normal nodes

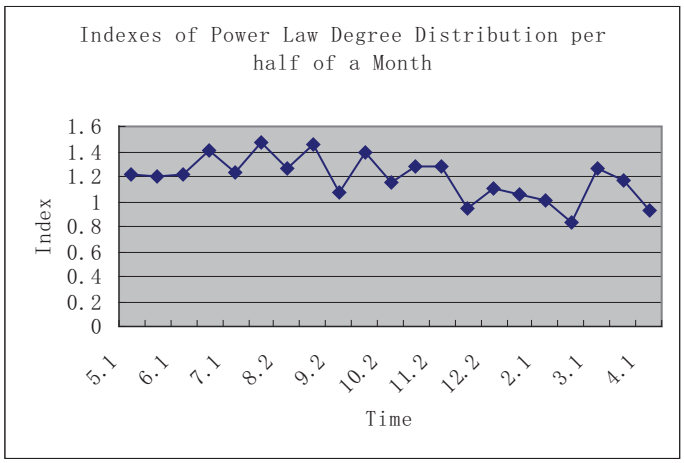
They are all Rest nodes.

3.2 Degree Distribution and Small World Properties

The degree distribution obeys power law. It is impossible to show all of them here, (some examples can be seen in Figure 3.4 (1)-(3)). Figure 3.3 shows that the exponents of these distributions hold steady.

An interesting observation is that that even for the two graphs of similar power law indexes, their structures can be totally different. For example, we can see that in Table 3.1, the exponent of power law degree distribution of September 15-30 is 1.066, December 1-is 15 0.949, and April 1-15 is 0.935. However, in Figure 3.4 (1)-(3), topology structures of these three graphs are quite different: one is within two main clusters connected by a bridge, one is two separated clusters, and the other is an international connected graph.

Figure 3.3 Exponents of Power Law Degree Distribution over half of Month



The cluster coefficiency CC >> C_{random} and average distances are roughly equal to the L_{random}, the average degree k=5.8, a number too close to 1. As a result, this is not a small world, but only a sparse social network.

We argue that indicators such as degree distribution and clustering coefficient are weak at describing particular local networks because they are global variables representing distribution of resources and efficiency of spreading. Thus local and ego-centric variables are needed to further explore this phenomenon.

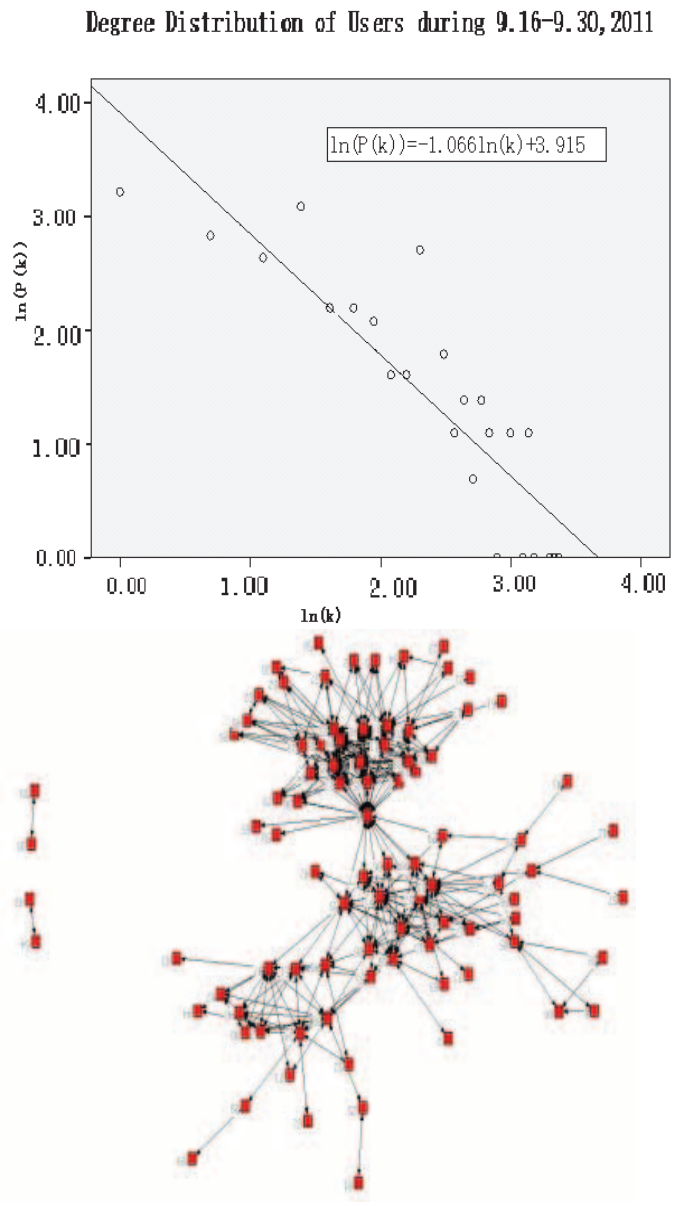


Figure 3.4 (1) The Power Law Degree Distribution and Topology Graphs of 15-30 Sep,2011

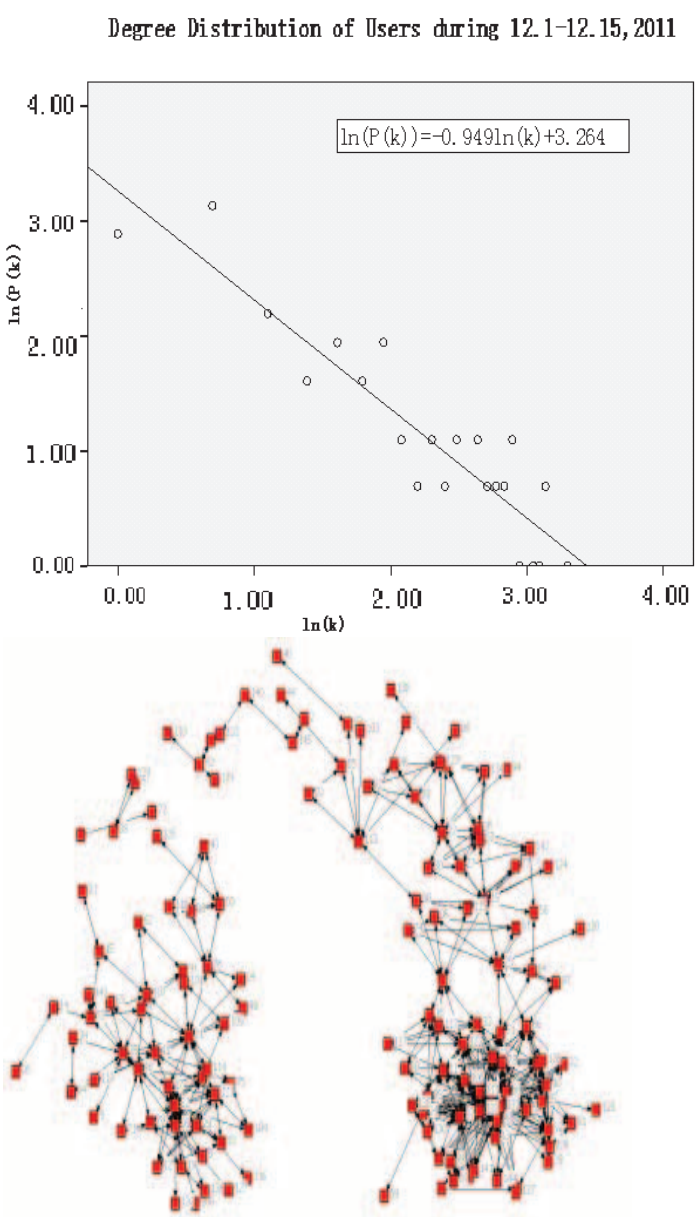


Figure 3.4 (2) The Power Law Degree Distribution and Topology Graphs of 1-15 Dec, 2011

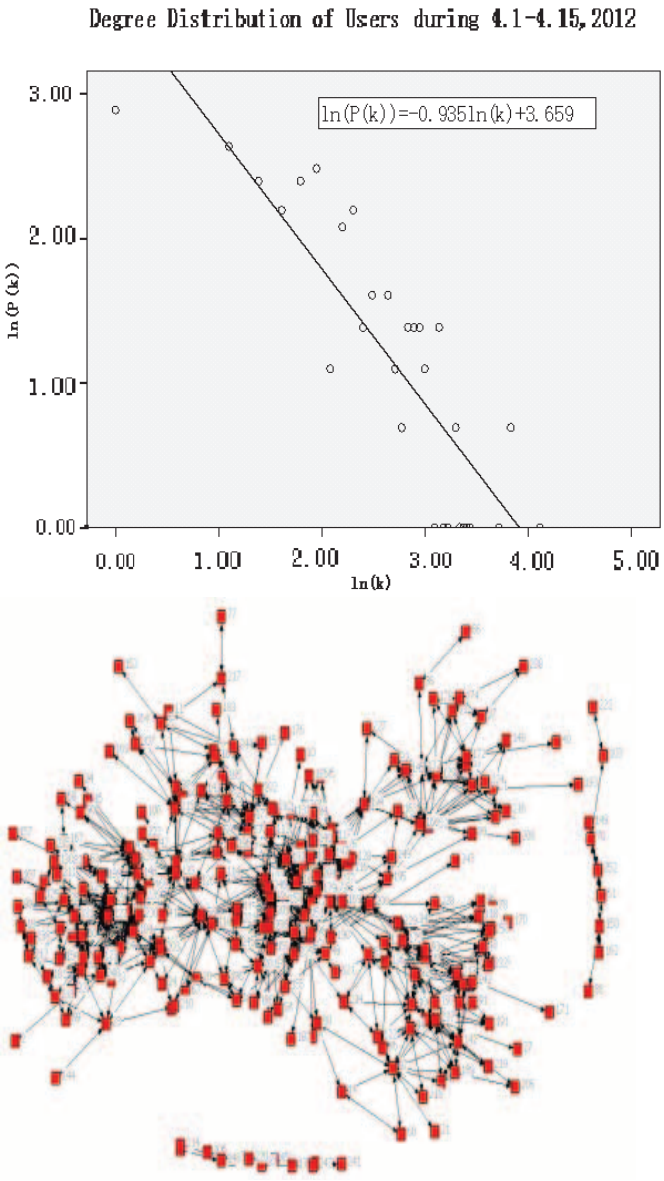


Figure 3.4 (3) The Power Law Degree Distribution and Topology Graphs of 1-15 April, 2012

3.3 Offline Interviews

3.3.1 Interactions Between Online and Offline

This is the mate selection process through online-offline interactions that we discovered from offline semi-interviews and participating observations. In figure 3.5, we can see that there are four kinds of offline relationship: Never meet again, couples, friends, and "419." Apparently, 419 (for one night→four one nine→419), which means one night based sex behavior, belongs to the High risk behavior that might spread HIV. In addition, for most of the 419 participants, they have no expectations or plans to meet again in the future (*given by interviewees 1,2,3,6,7,8*). The friend relationship means they define each other as friends to interact with, but not mates for conducting intimacy behaviors. There are some unwritten rules and regulations among friends, such as one cannot betray his friends by stealing their money, cheating them, or taking one's ex-boy-friend/s for himself. High frequency of 419 behavior is also harmful for one's reputation (*given by intervie*wees 2,3,7,8). As a result, the primarily platonic friendships are quite safe, which means there is few HRB except for a few 419 seekers who make friends for sex. Serious couples are independent from the two aforementioned relationships. Month/s long couples relationships are based on deep understanding and trust between each other, including health situations (*given by interviewees 4.5.6.7.8*). They thus keep HIV away from the relationship in this social arrangement. A couple's relationship based on 419 will not last long. *Interviewees 2 and 6* tell us the maximum duration of this type of relationship is 1 week. Consequently, the HRB mainly occurs between those in 419 relationships.

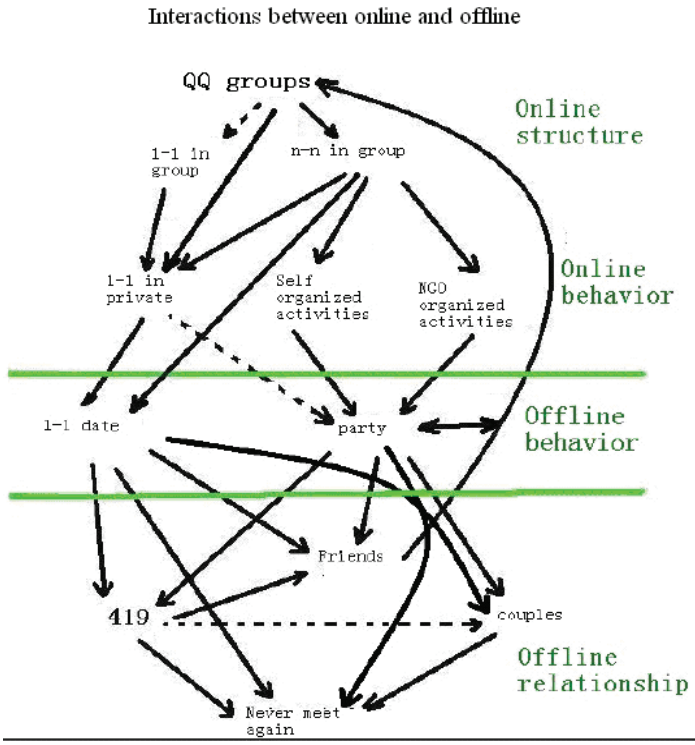


Figure 3.5 Interactions Between Online and Offline

There are two paths leading to 419 relationships. One path is from 1-1 date which means two people date to select a suitable mate similar to heterosexual relationships. The other path concerns parties invited to join in a 1-1 private conversations, which cannot be seen by others in the group, or held by both self organizations and NGOs. Although there are many more interactions in the figure 3.5, we focus our attention on 419 related paths for each role in structural positions.

3.3.2 Key Nodes

Figure 3.6 gives the paths to 419 among Key nodes. As they are the most popular people in the community, they are accosted every day by 1-1 private conversations. Besides, they have huge offline friendship networks which can lead them to join into various offline parties in bars and KTVs. This kind of party offers numerous opportunities to make new friends and who are searching for 419 relationships. Consequently key nodes frequently participate in HRB. Three other factors also play crucial roles in the decision to participate in HRB for key nodes:

1) Whether he is in couple relationship;

If one is in his couple relationship, he would have fewer opportunities for other relationships. The couples tend to strictly control each other to maintain their love life. So if they have HIV they may only spread HIV between themselves. But still they can continue frequent 419.

Having a boyfriend is also an important reason for sudden fading from online communities (*given by interviewees 1,2,3,5,8*). However, some interviewees report that they had frequently HRBs within their first week of breaking up (*given by interviewees 1,6,8*).

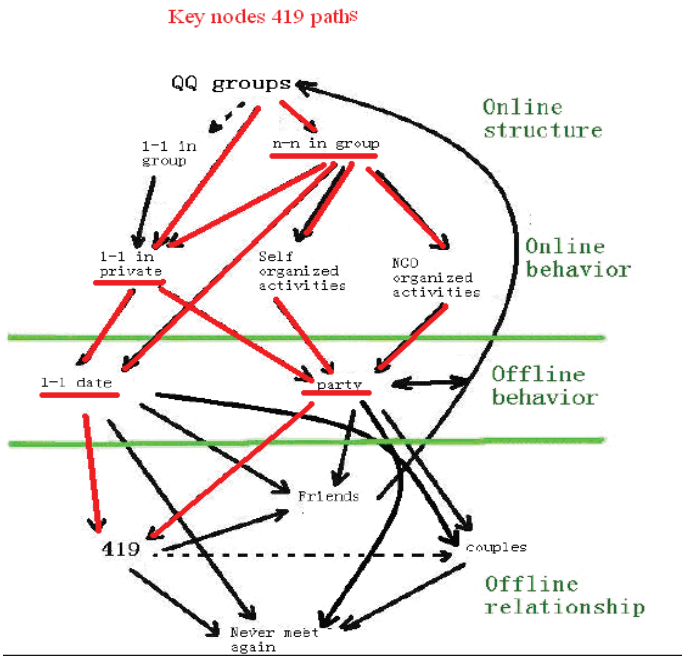


Figure 3.6 Key Nodes' Paths to 419

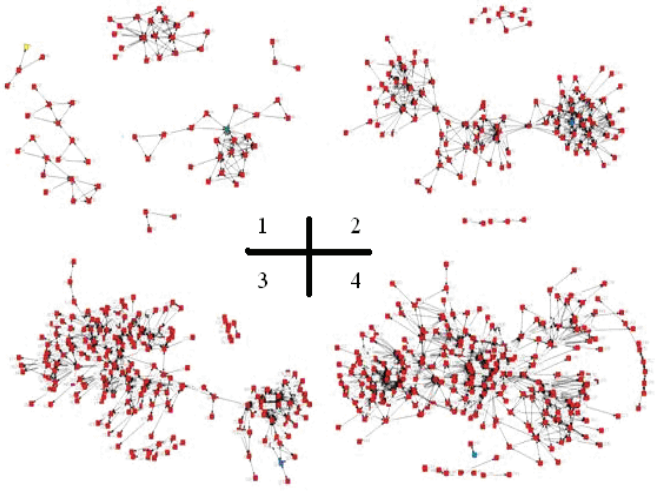


Figure 3.7 An example of key node's (dark green node in the graph) fading because of having new boyfriend (when he is the most popular node in the second graph). They are network graphs from Feb 2012 to March 2012.

2) Whether or not one has been a member or volunteer in NGO that works on HIV protection and Gay community cultivation.

An experience working in the NGO will help greatly them to understand the importance of condom usage and other HIV protecting strategies (*given by interviewees 1,6,7,8*). They thus have a deep impression and are more likely to keep themselves safe when they have sexual contact with another.

3) The years he has been a participant in this online community.

Another reason for the fading of online key nodes, gay homosexuals have their own psychological evolutions: In the first of the two years, a new participator is welcomed because he his "fresh." When it comes to the third and forth year, most of the new ones feel depressed then try start to stay away from the community. In the next two years he will understand that there are still some good in his friends and some love is worth believing in. After seven years, he will become active in the community again (*given by interviewees 2,3*).

3.3.3 Active participators

For active participants combining by both online and offline friends, they have more narrow paths to 419 relationship. If one has highly frequent HRBs, his reputation will suffer in these small groups. As a result, he will lose his friends who offer sufficient social sustainability and gay homosexual identification (*given by interviewees 1,3,6*). One thus can hardly build 419 relationships in parties without his all friends' absence. Nevertheless, 1-1 private conversations frequently happen in secret online that trigger HRBs because of activities in this role.

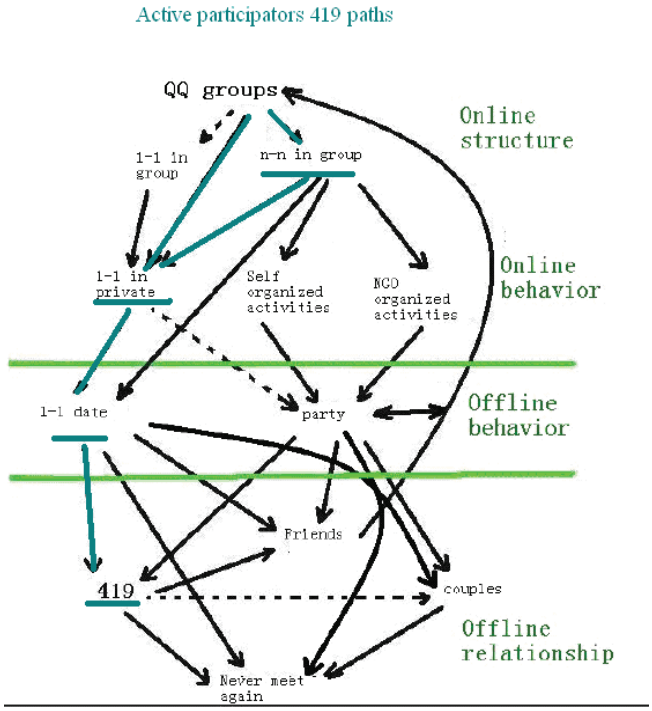


Figure 3.8 Active Participators' Paths to 419

3.3.4 Lurks and Bubbles

As the roles within the largest population, lurkers and bubbles have the fewest restrictions compared with all other roles. As a result, they can take various paths to 419 relationships. Only a few of them use n-n conversations in the QQ groups. They prefer,

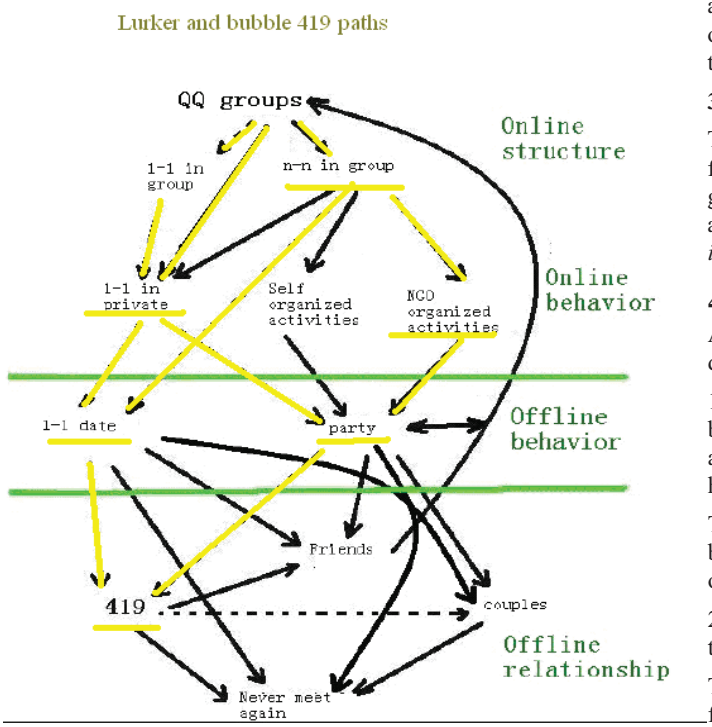


Figure 3.9 Lurker' and Bubble' paths to 419

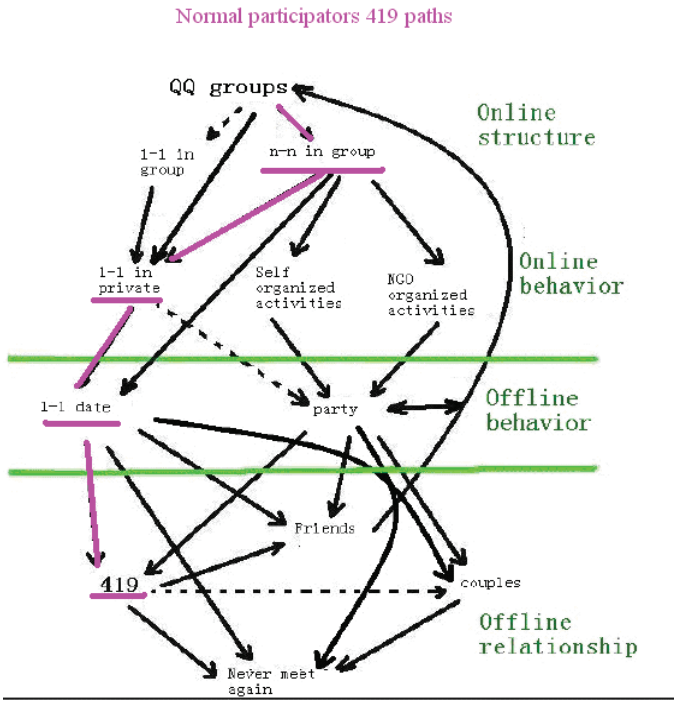


Figure 3.10 Normal nodes' Paths to 419

at least in the aspect of online behavior, 1-1 private chatting to find mates or joining in parties. In our participating observations, we found that a key node will handle more than 30 new messages in a day to apply for friendship and private conversation sent by lurkers and bubbles (*given by interviewees 2,6,7*). As well, it is astonishing that the interviewed lurker/bubble told us they examine every user in the groups in order to build connections and future relationships. One hopes, in this way, he will have opportunities to date or make friends with them. It is quite a large task, especially for new participators.

3.3.5 Normal Nodes

The rest of the users have their own individual degrees and frequency between lurkers and active participators in small groups. They thus have fewer restrictions (shown in figure 3.10) as well as fewer opportunities to accost and be accosted (*given by interviewee 2*).

4. Probable Hypothesis

As we lack of a larger random sampling, these results can only derive a hypothesis from our case study.

1) If a lurker who was key node (the node has max degree or the bridge node) a long time ago, had a sharp conversation frequency, and chooses to then lurk again, the probability of HRB is quite high.

This was disclosed when the key node interviewee had just broken up with his boyfriend. His frequency evolution can thus be observed as a sharp peek after a long time steady period at bottom.

2) The active nodes that suddenly disappear online will destroy the structure of the community, but have low probability of HRB.

The active nodes in small groups can be identified by a range of frequencies that get close to maximum. Concerned this also shows fading and returning processes are crucial and should be paid attention to.

3) The babbles and lurkers may have high level of HRB, especially for those who are still in their first two years in the community.

A large random offline sampling is needed to test this hypothesis.

4) Normal nodes (those who are not Key nodes, not in active groups, or are lurker & babble nodes) have low levels of HRB.

5. Discussions

As shown above, we first classify users into four roles. Second, we calculate online properties to show global factors are weak in predicting local HRBs. Third, from offline interviews and participating observations we focus on how the paths to HRB though online-offline interactions. Finally, based on individual features, multiple online structural positions, and changes in behavior patterns, we give four probable hypotheses to predict offline HRBs from online structures and individual features. This is called the ego-centric dynamical network analysis.

This work and hypothesis are based on limited fond and one-year data samples. Moreover, it is impossible to list all graphs and interview materials here. We have tried our best to present a fair representation in both field work and this analysis by providing complementary data analysis-interview log explanations. Though we have given our best effort on connecting with interviewees, our list is by most standards "thin" thus more extensive field work is needed to provide a more comprehensive picture.

Future research will include testing the hypothesis and providing key factors to help in the prediction of offline HRB. Comparison with theories calculating local components will also be conducted. Finally, evolution processes of friends surrounding target nodes will also be calculated.

6. REFERENCES

[1] Barabasi A.-L. and Albert R. 1999. Emergence of scaling in random networks. Science, 286:509–512.

[2] Barabási A L. 2003. Linked: The New Science of Networks. American Journal of Physics. Volume 71, Issue 4, 409

[3] Dodds P. S., Muhamad R. and Watts D. J. 2003. An experimental study of search in global social networks. Science, 301:827–829.

[4] Ravi K., Novak J., Tomkins A. 2006. Structure and evolution of online social networks in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 611-617.

[5] Wilson B. 2006. Ethnography, the Internet, and Youth Culture: Strategies for Examining Social Resistance and "Online-Offline" Relationships. Canadian Journal of Education, Vol. 29, 307-328.

[6] Khaled, R., Barr, P., Noble, J. and Biddle. R. 2006. Investigating social software as persuasive technology. Proc. Persuasive 2006, Springer, 104-107.

[7] Ploderer B., Howard S., Thomas P. 2008. Being online, living offline: the influence of social ties over the appropriation of social network sites. In Proceedings of ACM Conference on Computer Supported Cooperative Work 2008.

[8] Babbie E. 2010. The Basics of Social Research. Wadsworth Publishing.

[9] Jones S. et al. 1997. Virtual culture: identity and communication in cybersociety. Trowbridge: Cromwell Press 1997.

[10] Campbell J.E. 2004. Getting it online, 2004, The Harrington Park Press NY 13904-1580.

[11] Goldman A. Winter 2010. "The Heart of the Matter: Online or off, couples still have to click." California Magazine. http://alumni.berkeley.edu/news/california-magazine/winter-2010-inside-out/heart-matter. Retrieved 2010-12-28.

[12] Burt R. 1992. Structural Holes: The Social Structure of Competition. Cambridge: Harvard University Press.