You Are Who You Know: Inferring User Profiles in Online Social Networks

A. Mislove, B. Viswanath, K.P. Gummadi, P. Druschel
Presented by Xiannong Meng
For ENGR 139, Spring 2010

Inferring User Profile From Social Network

- Question to answer: Is it possible to *infer* the missing attributes of a user from the attributes provided by other users in the network?
- The answer is "quite possible!"
- This paper presents the results of a study to show such possibility.

Outline of the Presentation

- Major findings of the research
- Result and analysis
- Methods of Study
- Conclusions

Major Findings

- Certain user attributes can be inferred from social network with high accuracy when given information on as little as 20 percent of the users.
- Users with common attributes are more likely to be friends and form dense communities.
- A method of inferring missing attributes of a user from that of others in the social network is proposed.

Background Information

- Large number of users form social network
- As of November 2009
  - MySpace: 275 million users
  - Facebook: 300 million users
  - Orkut: 67 million users
  - LinkedIn: 50 million users
- Other social networks share media contents
  - YouTube
  - Flickr
  - Picasa
- Yet more social networks share blogs
  - LiveJournal
  - BlogSpot

User Information as Attributes

- Many of these social networks ask their users to establish user identity with attributes such as name, gender, age, interests, and others.
- These pieces of information can be considered as *attributes* of a user.
- Many users don't fill out all attributes because
  - Either forgetting to do it
  - Or concern of privacy

Inferring Attributes From Social Networks

- This study indicates that some missing attributes can be inferred from a user's friends from social networks!

Data Collected (1)

- Two sets of data are collected for the study.
- Rice University data set
  - Almost 4,000 students and alum of Rice University from the Rice Facebook group.
  - Collected names of the users and their list of friends (names only) from the Facebook group
  - Additional information such as matriculation year, graduation year, residential college, and major(s) or department are collected from other resources.

Data Collected (2)

- New Orlean data set
  - About 63,000 users in the New Orleans Facebook regional network
  - All attributes are collected in the process of crawling user profiles
    - Educational information
    - Tastes
    - Geographical information
    - Employers
    - And others

New Orleans Data Set (1)

- We created Facebook accounts on the New Orleans Facebook group, crawling information in parallel
- Discovered 90,269 users connected by 1,823,331 undirected links, for an average degree of 40.39
- We were able to download profiles of 63,731 users. This set is used for the study

New Orleans Data Set (2)

- Of the 63,731 user profiles, the right table shows what information is revealed in the profiles.

Attribute	% revealed
high school	68.9%
university	58.3%
employer	42.3%
interests	35.5%
location	19.3%

Friends with Common Attributes

- Two measures
  - $S_a = |\{(i,j) \text{ in } E, a_i = a_j\}| / |E|$
  - The number of links which connect two nodes with the same attributes divided by the total number of edges
  - $E_a = \text{sum}(T_i(T_i-1)) / (|U|(|U|-1))$
  - The summation is over all possible attributes for a node (user) and $|U| = \text{sum}(T_i)$
- The measure affinity is defined as A = S / E

Affinity Values for Various Communities

Users	Attributes	Affinity
Rice UG	College	4.49
	Major	2.33
	Year	1.97
Rice Graduate	Department	9.71
	School	4.02
	Year	1.79
New Orleans	High school	53.2
	Hometown	2.87
	Political views	1.86

Attribute Based Communities

- Given that we have observed a correlation between user attributes and links, it is natural to see if users who share similar attributes form communities, or dense clusters.
- Modularity is a measure between -1 and 1 where 0 means neutral and positive value means significant community and negative value means less community

The Modularity Measure

- Partition a social network into k community
- Let e be a symmetric k by k matrix whose element e(i,j) is the fraction of edges in the network that connect i and j
- a(i) = sum e(i,j) over j is the fraction of edges that touch vertices in community i
- Trace e = sum e(i,i) gives the fraction of edges in the network within the same community
- Modularity Q = sum [e(i,i) – a(i)*a(i)]

15

Modularity Among Different Communities
(Rice Undergraduates)

Attributes	Communities	Modularity
College, major, year	582	0.023
College, major	317	0.029
Year, major	147	0.045
major	52	0.055
College, year	44	0.248
year	7	0.259
college	9	0.384

Modularity Among Different Communities
(Rice Graduates)

Attributes	Communities	Modularity
Year	10	0.185
Department, school, year	124	0.292
Department, year	124	0.292
School, year	43	0.299
School	7	0.581
Department, school	28	0.587
Department	28	0.587

Inferring Attributes Globally

- First detect community at a global level
- Then compute the similarity among different communities
- With similar communities, similar attributes are detected
- See Figure 1 on page 256
- With about 20 percent of users reveal their information, the accuracy can be as high as 80 percent!

Determine the Global Community

- Removing edges until the network is partitioned
- To decide which edges to remove, a metric known as *betweeness centrality* is computed
  - Shortest paths between all pairs are computed
  - The number of shortest paths going over a particular edge is the measure of centrality of that edge
  - These edges become the connection between communities
  - When these edges are removed, the communities are revealed

19

References

- Mislove, A., Viswanath, B., Gummadi, & K.P., Druschel, P. (2010). You Are Who You Know: Inferring User Profiles in Online Social Networks. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining* (Brooklyn, NY, U.S.A., Feb 2010), 251-260. Accessed February 10, 2010 from: http://www.wsdm-conference.org/2010/proceedings/docs/p251.pdf