THE 18TH ACM SIGKDD CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING
KDD 2012 BEIJING
Beijing, China
August 12-16, 2012

User-Interest based Community Extraction in Social Networks

Diana Palsetia[‡], Md. Mostofa Ali Patwary[‡], Kunpeng Zhang, Kathy Lee, Christopher Moran, Yves Xie, Daniel Honbo, Ankit Agrawal, Wei-keng Liao, Alok Choudhary
Dept. of Electrical Engineering and Computer Science, Northwestern University
Evanston, IL 60208, USA
[‡]Corresponding authors:{drp925, mpatwary}@eecs.northwestern.edu

Introduction

A community is a densely connected subset of nodes(users) that is only sparsely linked to the remaining network. In this paper we focus in finding communities in social network based on user-generated content such as comments and tweets of millions of users in Facebook and Twitter respectively.

Data Model

The user generated content is used to formulate the ***dynamic*** network (graph). The nodes in the graph are Facebook walls or Twitter profiles (users). The edges between any two walls/profiles are weighted by *Jaccard index* (similarity coefficient):

$$w[i,j] = \frac{M[i,j]}{M[i,i]+M[j,j]+M[i,j]}$$

M[i,i] = unique users(uu) for wall/profile i

M[i,j] = common users(cu) between i and j

Table 1: Structural properties of the dataset

	Facebook	Twitter
Max(uu)	766,700	173,100
Avg(uu)	14,070	6,946
Avg(cu)	10	46
Total(uu)	22,795,352	2,215,581

Related Work

The widely used community detection algorithms are based on optimizing a metric, known as *modularity*(*Q*). These algorithms are in general known as *Greedy Agglomerative* (GA).

$$Q(C) = \sum e_{ii} - a_i^2$$

The eq. above, e_{ii} is the fraction of edges fall within group *i* and a_i is the total fraction of all ends of edges that are attached to vertices in group *i*. According to [1], maximum modularity does not necessarily reflect that a network has community structure.

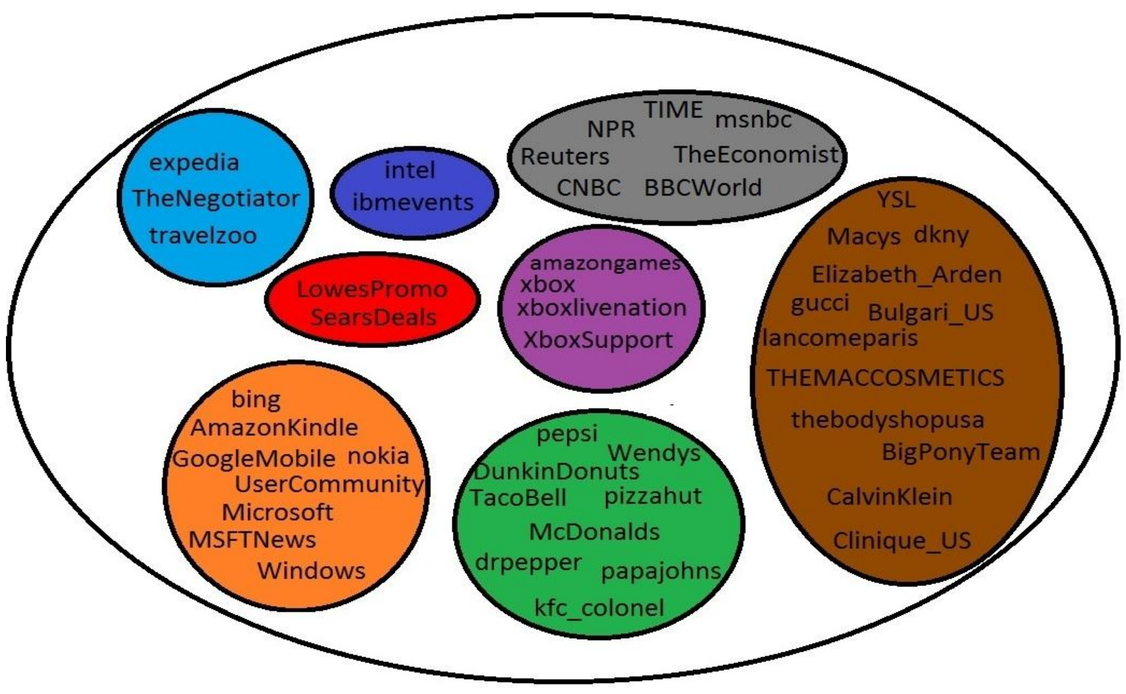

Figure 1: A large single community found by CNM [2]

Approach

Our Incremental Community Extraction(INC) algorithm:

- At each round, calls GA algorithm (specifically CNM but this can be changed)
- For each community that the GA algorithm outputs
  - INC either declares that as a final or
  - Recalls itself for that community to divide it further by discarding the influence of the communities identified in the previous steps

Results

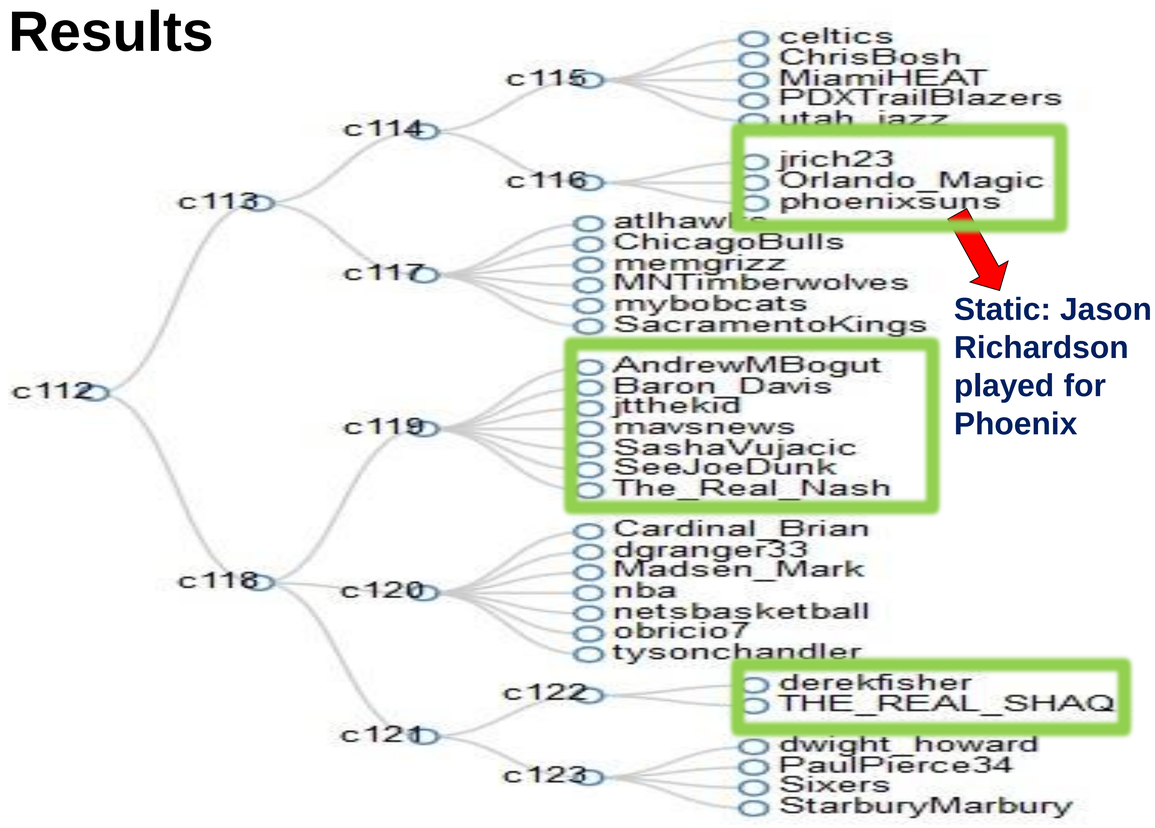

Figure 2: NBA Basketball (Static Twitter Dataset)

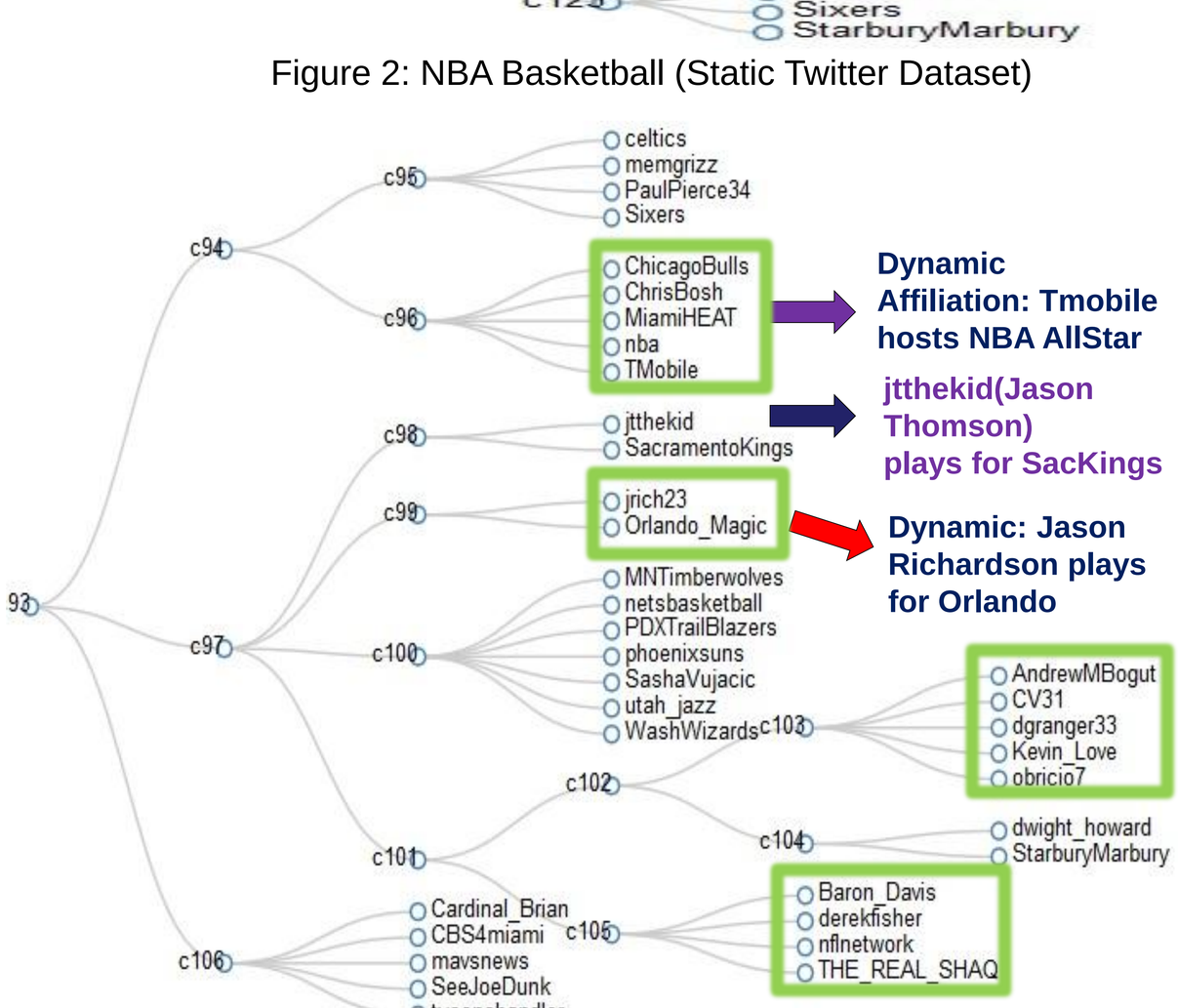

Figure 3: NBA Basketball (Dynamic Twitter Dataset)

Table 2: Comparison between CNM and INC

	CNM		INC	
	Q	Mod Den*	Q	Mod Den*
fb_dyn	0.11	2428.47	0.00	2622.88
tw_dyn	0.10	486.44	0.01	443.08
tw_stat	0.31	136.04	0.01	505.53

**Modularity density* is the sum over the clusters of the ratio between the difference of the internal and external degrees of the cluster and cluster size.

Conclusion

User-interest model finds affiliations that are constantly evolving either due to temporal or spatial activities. To overcome the limitations of the widely used modularity based algorithm (CNM), our approach incrementally extracts communities disregarding the influence of the communities identified in the previous steps. This allows us to extract better focused communities.

Selected References

[1] S. Fortunato and M. Barthelemy. Resolution limit in community detection. Proc. of National Academy of Sciences, 104(1):36, 2007.
[2] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. Phys. Rev. E, 70(6):066111, 2004.

Acknowledgment

This work is supported in part by NSF award numbers CCF-0621443, OCI-0724599, CCF-0833131, CNS-0830927, IIS-0905205, OCI-0956311, CCF-0938000, CCF-1043085, CCF-1029166, and OCI-1144061, and in part by DOE grants DE-FG02-08ER25848, DE-SC0001283, DE-SC0005309, DE-SC0005340, and DE-SC0007456.

Paper ID: SNAKDD18