A Diffusion of Innovation-Based Closeness Measure for Network Associations

Reihaneh Rabbany, Osmar R. Zaïane

Department of Computing Science
University of Alberta
Edmonton, Canada
rabbanyk,zaiane@ualberta.com

COMMPER Workshop on Mining Communities and People Recommenders
in conjunction with the IEEE International Conference on Data Mining (ICDM)
Vancouver, Canada
December 2011

How do we define a community?

Several definitions for communities in the network

- Density of edges within is greater than between
- Community membership based on a notion of similarity

Proximity measures based on feature vectors
e.g. Euclidean distance, Mahalanobis distance

iCloseness

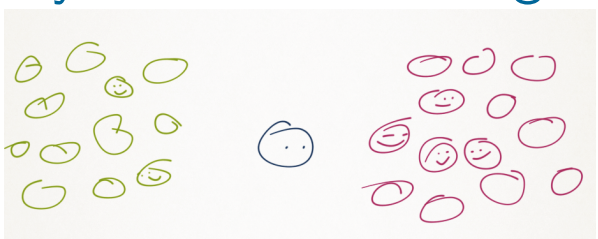
New closeness measure appropriate for mining communities

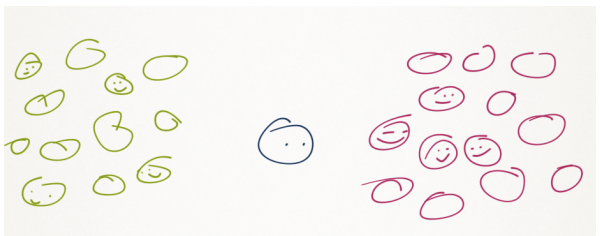
- Encapsulates the notion of membership in a community
- Reinforced by observations made on community dynamics in social networks with regard to the probability of joining a group based on the concept of Diffusion of Innovation (Backstrom *et al.* , 2006)

What is the basic idea of iCloseness?

likelihood of joining a community in social networks depends upon the number of pre-existing connections with group members and the density of edges between these members and other members in the group

i.e. if I am faced with two groups in which I already have friends and if I need to join one of these groups, I could choose either one but, there is a higher probability to join the group in which I have more friends. In addition if I am faced with two groups in which I have the same number of friends, there would be a higher probability that I would join the group in which the connectivity of my friends with the group is stronger





Outline

Addressing similar question as graph partitioning

- Spectral clustering method e.g. (Ng *et al.* , 2001)
  the sizes of the groups need to be fixed
- Cut minimization methods e.g. *ratio cut* (Chan *et al.* , 1993), *normalized cut* (Shi & Malik, 2000), and the *min-max cut* (Ding *et al.* , 2001)
  in favour of divisions into equal-sized parts

Community structure detection assumes that the networks divide naturally into some partitions and there is no reason that these partitions should be of the same size

Hierarchical clustering as the standard method

Hierarchical clustering: discovers natural divisions of social networks into groups, based on

- a metric measuring the similarity/closeness between vertices
- a quality function measuring the quality of a particular division (Chen *et al.* , 2009; Girvan & Newman, 2002; Guimera *et al.* , 2004; Newman & Girvan, 2004; Nicosia *et al.* , 2008)

  Modularity Q (Newman & Girvan, 2004; Clauset *et al.* , 2004), is a well-known example of such quality function in finding communities and serves as the basis of many other later proposed metrics

The iCloseness measure is based on the connection between theoretical models of diffusion in social networks to the community membership investigated by (Backstrom *et al.* , 2006)

Outline

Definition of iCloseness

Defined based on theory of *Diffusion of Innovations* – a theory of how, why, and at what rate new ideas and technology spread through cultures

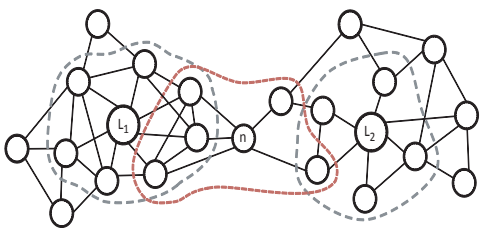
Specifically for the case of information networks, if we consider the act of joining a community as a behaviour that spreads within a network the idea of Diffusion of Innovations very naturally extends to community mining.

Socially, there is indeed advantage in joining a group that already includes friends that know each other and who are connected

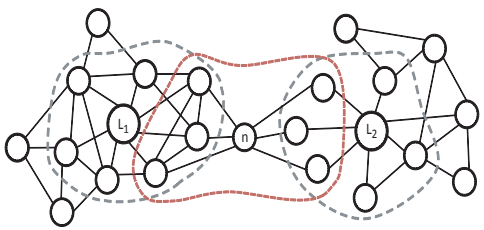
(Backstrom *et al.* , 2006)

Definition of iCloseness

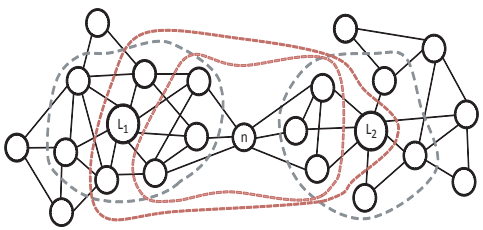
Measures closeness between two nodes and indicates how much these nodes tend to belong to the same community $\Rightarrow$ calculated based on their common neighbours



Intersection of neighbourhoods



Density of common neighbours



Expanding neighbourhoods

Node n should be assigned to community of leader L_1

Neighbours Scoring

$\aleph(v)$: neighbourhood of node v
i.e. $u \in \aleph(v)$ *iff* there exists a path of length at most δ to that node from v

- **neighbours scoring**: based on the depth of their neighbourhood and based on how dense they are connected

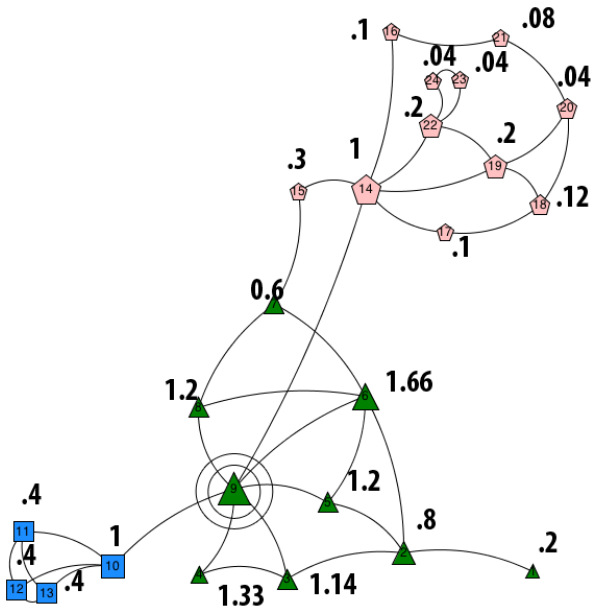
$$ns_1(u,v) = \begin{cases} 1 & \text{if } e(u,v) \in E \\ 0 & \text{otherwise} \end{cases}$$

$$E_1(v) = \{e(u,v) \in E\}$$

$$ns_l(u,v) = ns_{l-1}(u,v) + \sum_{e(u,m) \in E - E_{l-1}(v)} ns_{l-1}(m,v) \times \frac{e(u,m)}{\sum_i e(i,m)}$$

$$E_l(v) = E_{l-1}(v) \cup \{e(i,j) \in E \mid \exists\, e(j,k) \in E_{l-1}(v)\}$$

An example of scored neighbourhood



Closer nodes to node 9 have higher neighbouring scores; neighbours that are more densely connected to 9, also obtain higher scores

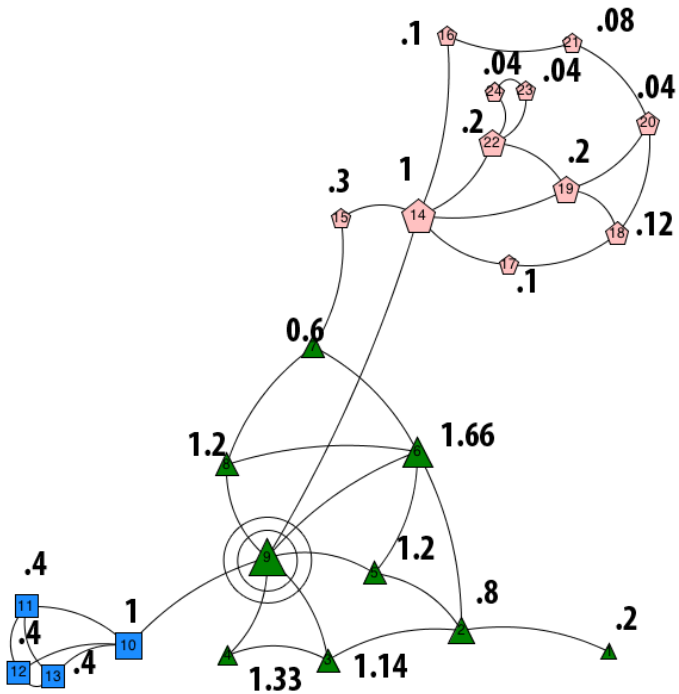
Computing iCloseness

The iCloseness of node v_1 and v_2, is obtained as:

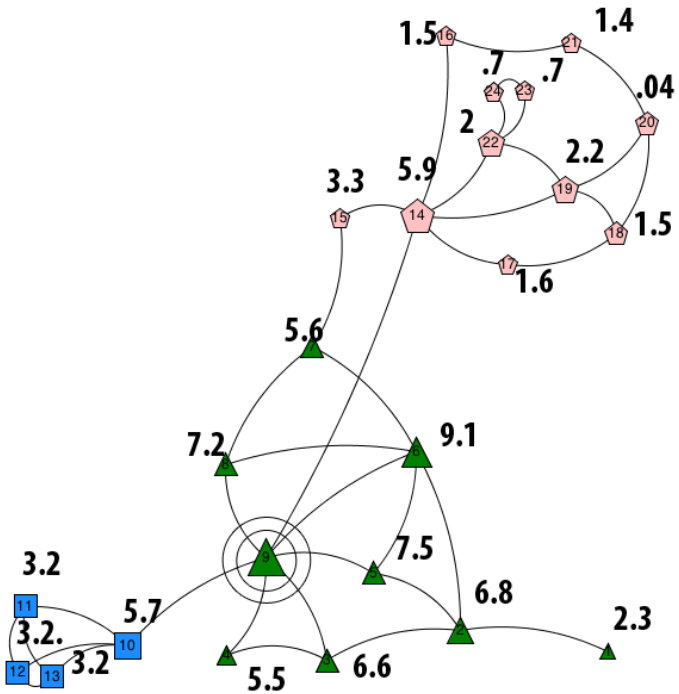
$$iCloseness(v_1, v_2) = \sum_{u \in \aleph(v_1) \cap \aleph(v_2)} ns(u, v_1) \times ns(u, v_2)$$

The neighbourhood threshold(δ) determines the maximum neighbourhood level which should be tuned based on the application e.g. a small number (*e.g.* 3) for social networks, because of the six-degree separation theory

Example revised: with iCloseness values



Scored neighbourhoods



Marked by iCloseness values

The same example but marked with iCloseness of all nodes to node 9

Applying iCloseness in community detection

Top Leaders first introduced in (Rabbany Khorasgani *et al.* , 2010)
Inspired by the well-known k-medoids clustering algorithm

Iteratively

- elects k representative nodes as leaders
- associats followers to one of these leaders to form communities based on the relations/links between nodes

We adopted the original Top Leaders algorithm to use iCloseness when determining the community memberships of nodes

iTop Leaders algorithm

initialize k leaders
repeat
  {finding communities}
  for all Node v $\in$ G **do**
    if v $\notin$ leaders **then**
      find community of v
    end if
  end for
  {updating leaders}
  for all $\ell \in$ leaders **do**
    $\ell \leftarrow \underset{v \in \text{Community}(\ell)}{\arg\max}$ Centrality(v)
  end for
until there is no change in the leaders

Leader
the most central member in its community

Community
a leader and the follower nodes associated with it

Association of followers to the *iClosest* leader

depth $\leftarrow 1$
CanList $\leftarrow$ leaders $\cap\ \aleph(\aleph(v))$
CanList $\leftarrow \underset{\substack{\ell \in \text{CandList} \wedge \\ \text{iCloseness}(v,\ell) > \gamma}}{\arg\max}\ \text{iCloseness}(v, \ell)$
if $|\text{CanList}| = 0$ **then** {No candidate leader}
  if Centrality(v) $< \epsilon$ **then** {Noise}
    associate v as an outlier
  else {Powerful but free}
    associate v as a hub
  end if
else if $|\text{CanList}| > 1$ **then** {Many candidates}
  associate v as a hub
else {Only one candidate leader in CanList}
  associate v to the only leader in CanList
end if

- the *iCloseness*
- the view
- speed
- k

Updating leaders

The election of the node with the highest centrality in each community

$$\ell \leftarrow \underset{v \in \mathit{Community}(\ell)}{\arg\max} \mathit{Centrality}(v)$$

The centrality of nodes in a community measures the relative importance/popularity of a node within that group

Initialization

Leaders are central nodes in their community $\Rightarrow$
the k most central nodes that none of them belong to the same community

Starts from the most central node, and adds the next central one to the current set of leaders only if it is not too *iClose* to any of the current leaders

A detailed comparison of different initialization methods for the Top Leaders is presented in (Rabbany Khorasgani *et al.* , 2010)

Outline

Evaluating community mining methods

1. **Well-known (small) real world data-sets**

   Comparing using a measures of agreement
   - Adjusted Rand Index (ARI) (Santos & Embrechts, 2009)
     $$[-1(noagreementatall)\ldots 1(fullagreement)]$$
   - Normalized Mutual Information (NMI) (Lancichinetti & Fortunato, 2009)
     $$[0(partitionsareindependent)\ldots 1(partitionsareidentical)]$$

   Karate and WKarate Club (weighted) (Zachary, 1977), Sawmill Strike (Pajek, n.d.), NCAA Football Bowl Subdivision (Xu *et al.* , 2007), and Politician Books (Krebs, n.d.)

2. **Synthesized with characteristics similar to real-world networks**

   LFR benchmarks (Lancichinetti *et al.* , 2008): Power-law distribution over degrees and community sizes & Mixing parameter μ: each node shares a fraction of its edges with the nodes of other communities

3. **Large real networks with no ground-truth**

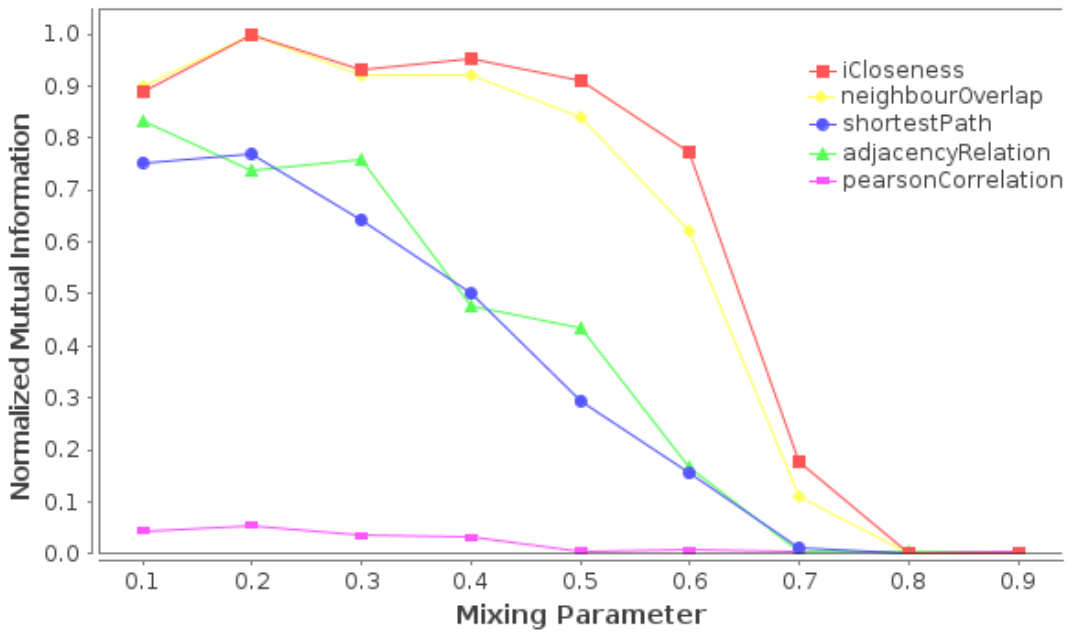
   Co-purchasing network of Amazon.com (Clauset *et al.* , 2004)

   $815,223$ nodes and $3,426,127$ edges

Results on real world benchmarks

dataset	measure	ARI	NMI
Karate (2 groups, 34 nodes, 78 edges)	Shortest Path	.669	.655
	Adjacency Relation	.771	.732
	Neighbour Overlap	.446	.383
	Pearson Correlation	.328	.324
	iCloseness	**1.**	**1.**
Strike (3 groups, 24 nodes, 38 edges)	Shortest Path	.935	.926
	Adjacency Relation	.903	.834
	Neighbour Overlap	.819	.763
	Pearson Correlation	.109	.307
	iCloseness	**1.**	**1.**
PolBooks (2 groups, 105 nodes, 441 edges)	Shortest Path	.647	.542
	Adjacency Relation	.630	.573
	Neighbour Overlap	.687	.585
	Pearson Correlation	.053	.157
	iCloseness	**.769**	**.696**
Football (11 groups, 180 nodes, 787 edges)	Shortest Path	.689	.559
	Adjacency Relation	.431	.753
	Neighbour Overlap	.970	.948
	Pearson Correlation	.082	.007
	iCloseness	**.996**	**0.989**

Comparing on synthesized benchmarks



Comparison of different measures on LFR synthesized benchmarks. The horizontal axis represent datasets with different mixing parameter μ, and the vertical axis is the accuracy. Different curves stand for different measures.

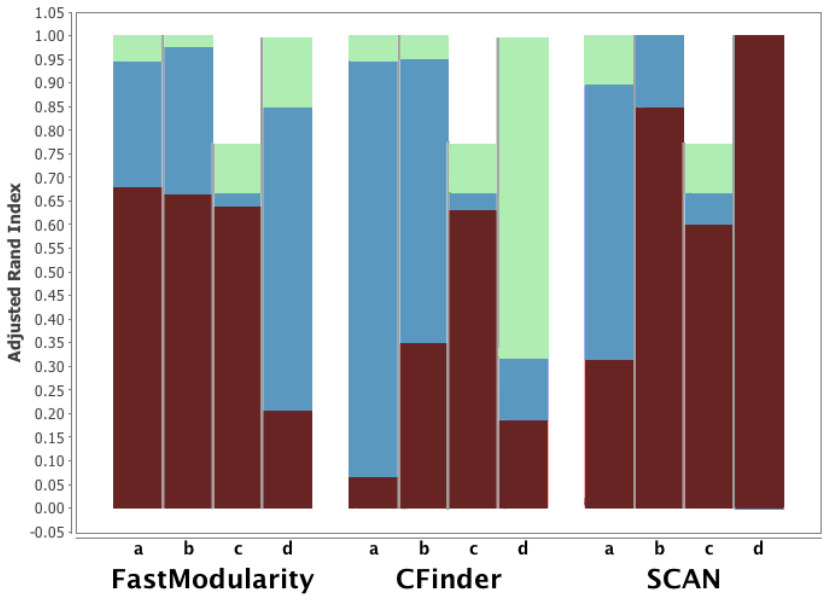
$$|V| = 1000, degree_{avg} = 20, degree_{max} = 50, \mu = 0 : .1 : 1, |C|_{min} = 20, |C|_{max} = 100$$

Comparing iTop Leaders with other algorithms

iTop Leaders equipped with iCloseness v.s. three well-known community mining approaches

1. FastModularity (Clauset *et al.* , 2004)
   Community: a particular partitioning with maximum quality compared to random
2. CFinder (Palla *et al.* , 2005)
   Community: union of adjacent cliques
3. SCAN (Xu *et al.* , 2007)
   Community: nodes that are structurally reachable

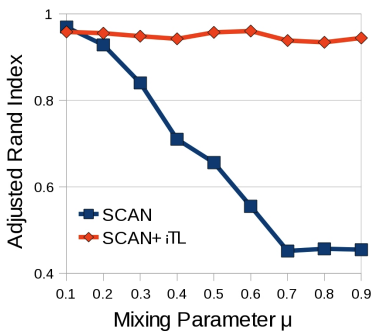
Results on real world benchmarks



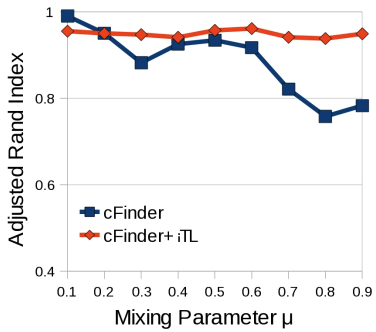
a) Karate, b) Strike, c) Politician Books and d) Football

- the red bar: accuracy of results obtained by opponent method
- the blue bar: result of iTop Leaders seeded with the same k
- the green bar: the ideal when iTop Leaders is seeded with correct k

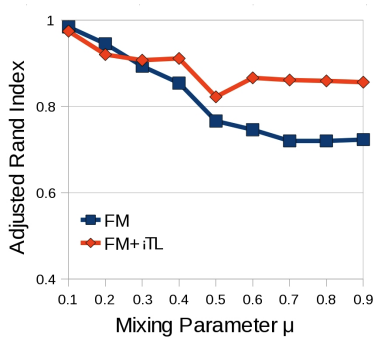
Results on synthesized benchmarks



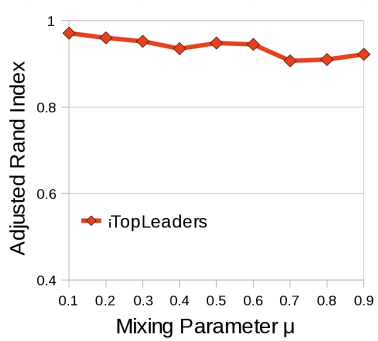
SCAN



CFinder



FastModularity



GroundTruth

The first three plots compare ARI of iTop Leaders (iTL) and other contenders as a function of mixing parameter μ. The last one shows the results of iTop Leaders given the correct k.

Results on large scale real-world data set

Amazon network

- ✖ CFinder: did not terminate successfully
- ✖ SCAN: did not terminate successfully
- ✔ FastModularity: x10 slower, $Q = .77$
- ✔ iTop Leaders with k_{FM}, $Q = 0.45$

When ground truth is not available, *modularity (Q)* is typically used to assess the quality of discovered communities

- $0.3 \geq$ shows a significantly good partitioning (Clauset *et al.* , 2004)
- higher modularity does not ensure a higher accuracy
- our algorithm detected 89865 hubs

  are not member of any specific community and this decreases the modularity significantly

Parameters

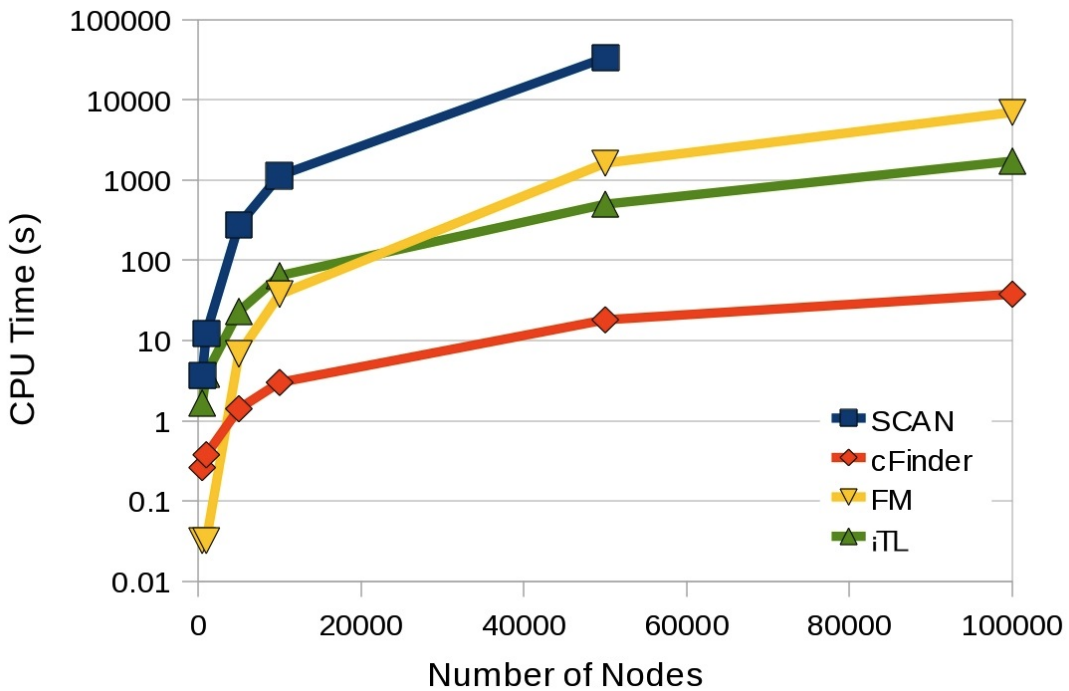
iCloseness

1. the neighbourhood threshold (δ)
   based on the application e.g. 3 for social networks

Top Leaders

1. the number of desired communities(k)
   Top leaders always improves the results with same k
   While k is mostly not correct or even close
   e.g. in the synthesized benchmarks with $33\pm5 \Rightarrow$
   FastModularity: 12 ± 6, CFinder:1182 ± 464, and Scan: 299 ± 127
2. the outlier threshold (γ), and the hub threshold (ϵ)
   Control the characteristics of the final output
   If both zero $\Rightarrow$ no outliers and disjoint clusters

Running time of iTop Leaders v.s. other methods



The running time of iTop Leaders, FastModularity, CFinder and SCAN

Conclusions

- Intersection Closeness to assess the proximity of a node to a community representative
- Based on the theory of diffusion of innovation
  which states that the probability of joining a group depends on the number of existing friends in the group and their connectedness
- The experimental results of Top Leaders algorithm equipped with *iCloseness* show high accuracy and effectiveness
  in both real and synthesized networks, compared against commonly used closeness/distance measures as well as the state-of-the-art community mining methods

- Questions?

Backstrom, Lars, Huttenlocher, Dan, Kleinberg, Jon, & Lan, Xiangyang. 2006.
Group formation in large social networks: membership, growth, and evolution.
In: ACM SIGKDD.

Chan, P.K., Schlag, M. D. F., & Zien, J. Y. 1993.
Spectral k-way ratio-cut partitioning and clustering.
Pages 749–754 of: Proceedings of the 30th International Conference on Design Automation.

Chen, J., Zaïane, O. R., & Goebel, R. 2009.
Detecting Communities in Social Networks using Max-Min Modularity.
In: SIAM Data Mining Conference.

Clauset, A., Newman, M. E. J., & Moore, C. 2004.
Finding community structure in very large networks.
Phys. Rev. E, **70**, 066111.

Ding, C., He, X., Zha, H., Gu, M., & Simon, H. D. 2001.
A Min-max Cut Algorithm for Graph Partitioning and Data Clustering.
Pages 107–114 of: ICDM '01: Proceedings of the 2001 IEEE International Conference on Data Mining.

Girvan, M., & Newman, M. E. J. 2002.
Community structure in social and biological networks.
Proceedings of the National Academy of Sciences of the USA.

Guimera, R., Sales-pardo, M., & Amaral, L. A. N. 2004.
Modularity from fluctuations in random graphs and complex networks.
Phys. Rev. E, **70**, 025101.

Krebs, Valdis.
Books about US politics.
http://www.orgnet.com/.

Lancichinetti, Andrea, & Fortunato, Santo. 2009.
Community detection algorithms: A comparative analysis.
Phys. Rev. E, **80**(5), 056117.

Lancichinetti, Andrea, Fortunato, Santo, & Radicchi, Filippo. 2008.
Benchmark graphs for testing community detection algorithms.
Physical Review E, **78**, 046110.

Newman, M. E. J., & Girvan, M. 2004.
Finding and evaluating community structure in networks.
Physical Review E, **69**.

Ng, A., Jordan, M., & Weiss, Y. 2001.
On spectral clustering: Analysis and an algorithm.
Pages 849–856 of: NIPS.

Nicosia, V., Mangioni, G., Carchiolo, V., & Malgeri, M. 2008.
Extending modularity definition for directed graphs with overlapping communities.

Pajek.
http://vlado.fmf.uni-lj.si/pub/networks/pajek/.

Palla, G., Derenyi, I., Farkas, I., & Vicsek, T. 2005.
Uncovering the overlapping community structure of complex networks in nature and society.
Nature, **435**, 814–818.

Rabbany Khorasgani, Reihaneh, Chen, Jiyang, & Zaiane, Osmar R. 2010.
Top leaders Community Detection Approach in Information Networks.
In: 4th SNA-KDD Workshop on Social Network Mining and Analysis.

Santos, Jorge M., & Embrechts, Mark. 2009.
On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification.
Pages 175–184 of: ICANN (2).

Shi, J., & Malik, J. 2000.
Normalized cuts and image segmentation.
IEEE. Trans. on Pattern Analysis and Machine Intelligence.

Xu, X., Yuruk, N., Feng, Z., & Schweiger, T. A. J. 2007.
SCAN: a structural clustering algorithm for networks.
Pages 824–833 of: ACM SIGKDD.

Zachary, W. W. 1977.
An information flow model for conflict and fission in small groups.
Journal of Anthropological Research, **33**, 452–473.