Meerkat

On the Practice of Evaluation for Community Mining in the Presence of Attributes

Reihaneh Rabbany and Osmar R. Zaïane

Department of Computing Science
University of Alberta
Edmonton, Alberta, Canada

Workshop on Multiplex & Attributed Network Mining @ ASONAM’2015 – Paris, August 25, 2015

University of Alberta - Edmonton

Edmonton, capital of Alberta, is the <u>5th largest city</u> in Canada with more than 1 million people.
The University of Alberta is the second largest university in the country in terms of research funding

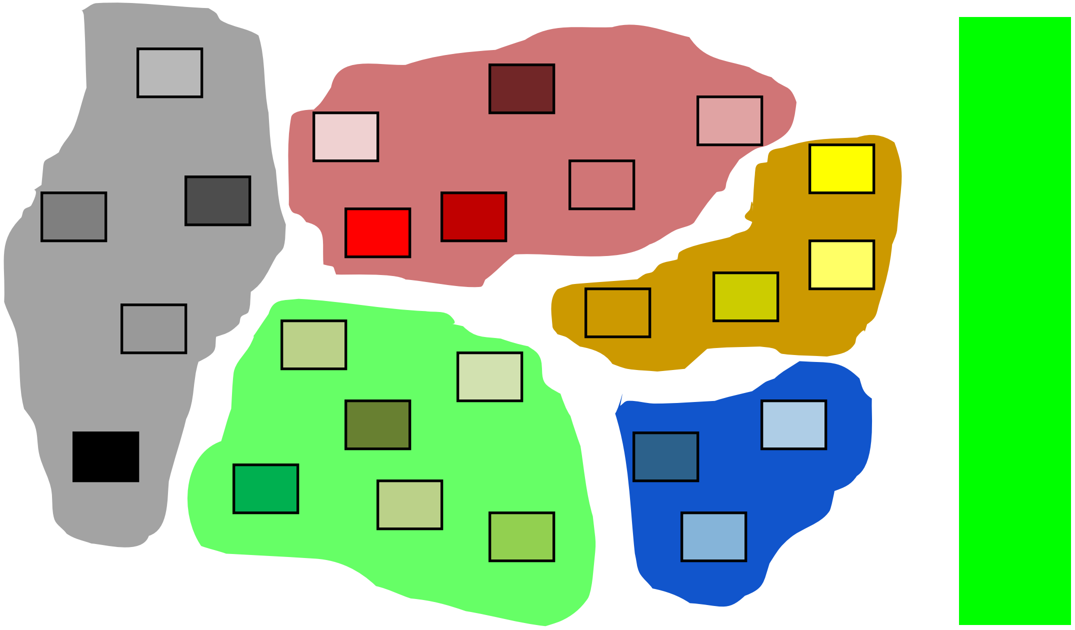
On the Practice of **Evaluation** for **Community Mining** in the Presence of **Attributes**

1- Community Mining

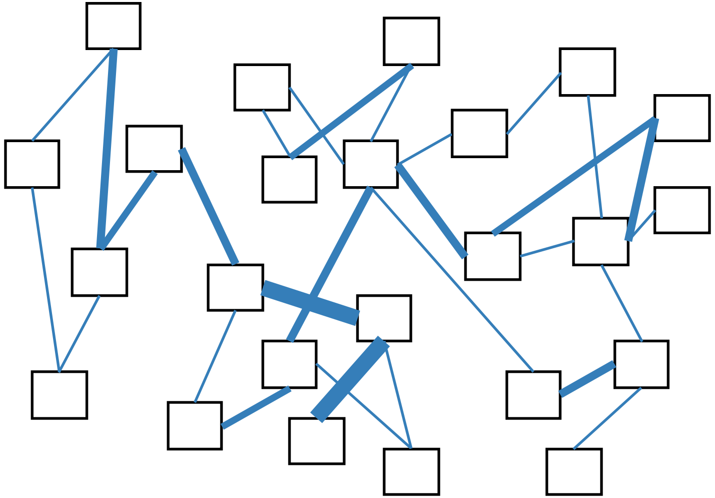
2- Validation of Community Mining

3- Suggest the use of Attributes in Community Mining

Clustering: The process of putting *similar* data points together.

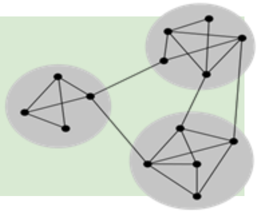


Clustering,
Grouping,
Partitioning data
based on attribute values



How to partition a graph of (attributed) nodes?

Modular Structure of Networks



One fundamental property of real networks

- Application such as module identification in biological networks
  - Protein-protein interaction networks outline protein complexes and parts of pathways
- Intermediate step for further analyses of networks such as link and attribute prediction
  - For example clusters of hyperlinks between web pages in the WWW outline pages with closely related topics, and are used to refine the search results

Motivating Example

Hypothetical telecom data

	Name	Phone Number	City	Plan	Avg. 3m Profit
1	John Smith	647 225 8085	Toronto	2y	($12)
3	John Simon	780 886 5053	Edmonton	3y	$189.45
4	Randy Regal	705 234 6767	Toronto	3y	$77.10
6	Mary Tasear Smith	780 334 3434	Edmonton	3y	$369.00
7	Susan Willcox	780 291 6063	Edmonton	2y	$131.00
8	Martha Witherby	780 322 9768	Edmonton	3y	$459.37
11	Kurt Locke	780 654 1121	Edmonton	3y	$830.00
12	Kent Wafegert	647 631 0348	Toronto	3y	$38.78
15	Brent Mavka	403 566 7372	Calgary	2y	$299.29
17	Wayne Jones	780 236 3006	Edmonton	3y	$236.06
18	Patty Klien	780 550 1819	Edmonton	1y	$50.18
20	Morris Slevchuk	780 434 6280	Edmonton	3y	$628.01
21	Patrick Klum	403 337 9291	Calgary	3y	$33.79
22	Wilma Renton	780 118 2388	Edmonton	3y	$8.00
24	Ben Rikon	403 262 3134	Calgary	3y	($26.23)
26	Maggie Wong	226 882 0911	Toronto	2y	$89.11
28	Karen Pollonts	403 750 9201	Calgary	3y	$92.75
31	Monica Kwalshuck	403 210 4448	Calgary	3y	$1,044.48
33	Natalie May	403 409 6223	Calgary	3y	$0.96

ID	Name	Phone Number	City	Plan	Avg. 3m Profit
24	Ben Rikon	403 262 3134	Calgary	3y	($26.23)
1	John Smith	647 225 8085	Toronto	2y	($12)
33	Natalie May	403 409 6223	Calgary	3y	$0.96
22	Wilma Renton	780 118 2388	Edmonton	3y	$8.00
21	Patrick Klum	403 337 9291	Calgary	3y	$33.79
12	Kent Wafegert	647 631 0348	Toronto	3y	$38.78
18	Patty Klien	780 550 1819	Edmonton	1y	$50.18
4	Randy Regal	705 234 6767	Toronto	3y	$77.10
26	Maggie Wong	226 882 0911	Toronto	2y	$89.11
28	Karen Pollonts	403 750 9201	Calgary	3y	$92.75
7	Susan Willcox	780 291 6063	Edmonton	2y	$131.00
3	John Simon	780 886 5053	Edmonton	3y	$189.45
17	Wayne Jones	780 236 3006	Edmonton	3y	$236.06
15	Brent Mavka	403 566 7372	Calgary	2y	$299.29
6	Mary Tasear Smith	780 334 3434	Edmonton	3y	$369.00
8	Martha Witherby	780 322 9768	Edmonton	3y	$459.37
20	Morris Slevchuk	780 434 6280	Edmonton	3y	$628.01
11	Kurt Locke	780 654 1121	Edmonton	3y	$830.00
31	Monica Kwalshuck	403 210 4448	Calgary	3y	$1,044.48

Not enough profit

Plan	Avg. 3m Profit
3y	($26.23)
2y	($12)
3y	$0.96
3y	$8.00
3y	$33.79
3y	$38.78
1y	$50.18

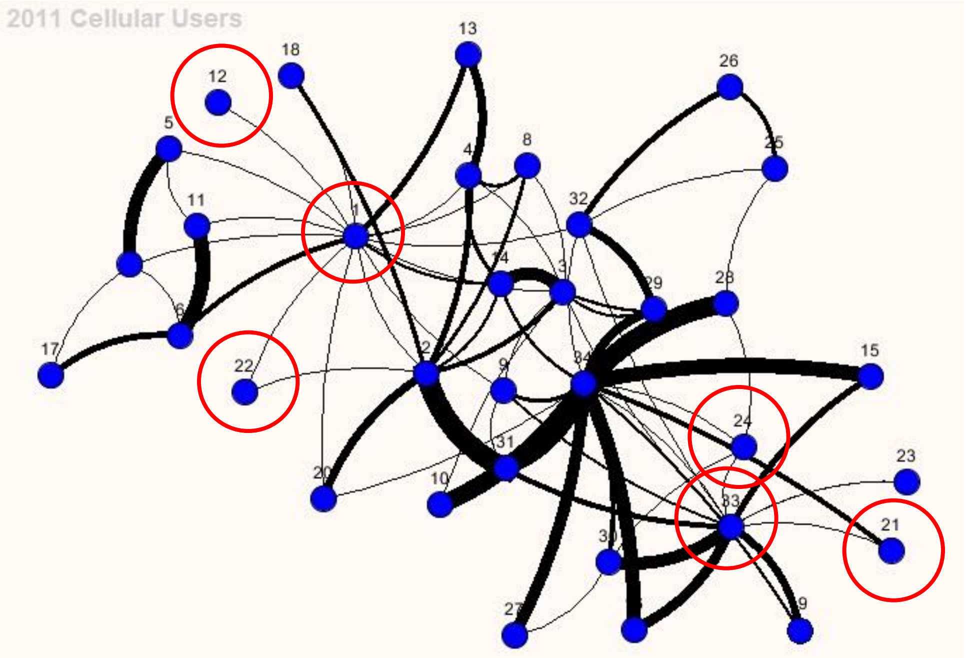
Assumption: Customers are **i**ndependent
Values are **i**dentically **d**istributed

6 least profitable customers
Could be the wrong decision

19 customers up for plan renewal
Which one to renew?
Which one to give incentive to stay?

Sort by profit in the last 3 months
Do not renew or give incentive if profit < $50 (?)

ID	Name	Phone Number	City	Plan	Avg. 3m Profit
24	**Ben Rikon**	**403 262 3134**	**Calgary**	**3y**	**($26.23)**
1	**John Smith**	**647 225 8085**	**Toronto**	**2y**	**($12)**
33	**Natalie May**	**403 409 6223**	**Calgary**	**3y**	**$0.96**
22	**Wilma Renton**	**780 118 2388**	**Edmonton**	**3y**	**$8.00**
21	**Patrick Klum**	**403 337 9291**	**Calgary**	**3y**	**$33.79**
12	**Kent Wafegert**	**647 631 0348**	**Toronto**	**3y**	**$38.78**
18	Patty Klien	780 550 1819	Edmonton	1y	$50.18
34	Aly Huffington	403 255 0304	Calgary	3y	$55.03
29	Iris Cristle	403 644 1423	Calgary	3y	$64.14
32	Fred Couros	416 773 2234	Toronto	3y	$73.22
23	Ryan Waters	403 715 7550	Calgary	3y	$75.550
4	Randy Regal	705 234 6767	Toronto	3y	$77.10
30	Gunther Twallaby	403 778 6040	Calgary	3y	$78.31
26	Maggie Wong	226 882 0911	Toronto	2y	$89.11
25	Jun Liu	226 690 4241	Toronto	3y	$90.42
9	Wanda Rhymes	403 441 2534	Calgary	3y	$92.00
28	Karen Pollonts	403 750 9201	Calgary	3y	$92.75
7	Susan Willcox	780 291 6063	Edmonton	2y	$131.00
3	John Simon	780 886 5053	Edmonton	3y	$189.45
17	Wayne Jones	780 236 3006	Edmonton	3y	$236.06
15	Brent Mavka	403 566 7372	Calgary	2y	$299.29
6	Mary Tasear Smith	780 334 3434	Edmonton	3y	$369.00
16	Brian Olso	403 939 7574	Calgary	3y	$430.78
8	Martha Witherby	780 322 9768	Edmonton	3y	$459.37
14	Kim Cho	780 434 2399	Edmonton	3y	$542.00
20	Morris Slevchuk	780 434 6280	Edmonton	3y	$628.01
5	Jane Smith	780 233 5645	Edmonton	2y	$673.38
2	Joe Burns	416 345 6060	Toronto	3y	$724.00
19	Greg Aderan	403 332 7468	Calgary	3y	$746.82
13	Megan Potink	780 432 5623	Edmonton	3y	$802.00
11	Kurt Locke	780 654 1121	Edmonton	3y	$830.00
10	Julie Austinshaur	403 223 7654	Calgary	3y	$983.12
31	Monica Kwalshuck	403 210 4448	Calgary	3y	$1,044.48
27	Joe Garther	416 224 1109	Toronto	3y	$1,100.10



Inter-call network with call frequency

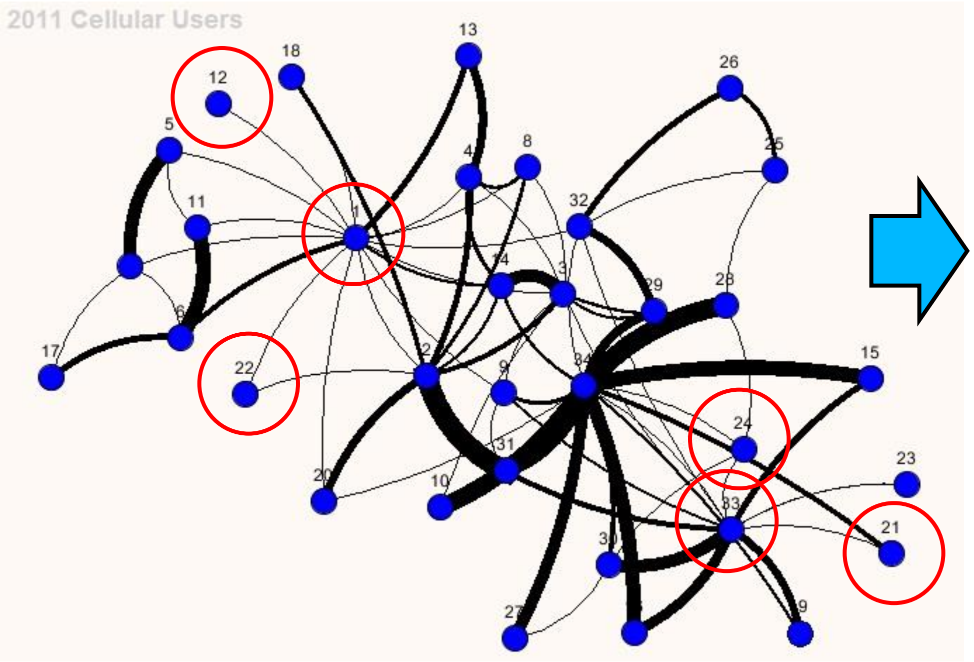
34 customers interconnected with the 19 to renew.

Which one to renew?

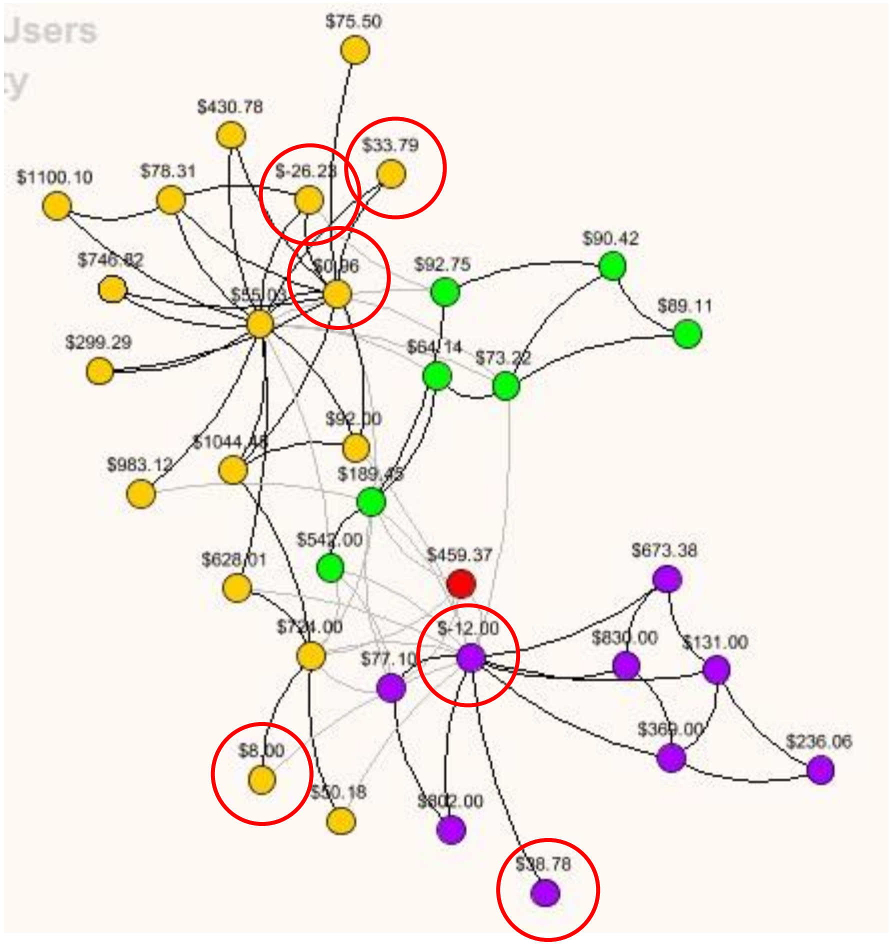
Which one to give incentive to stay?

Additional data was required:

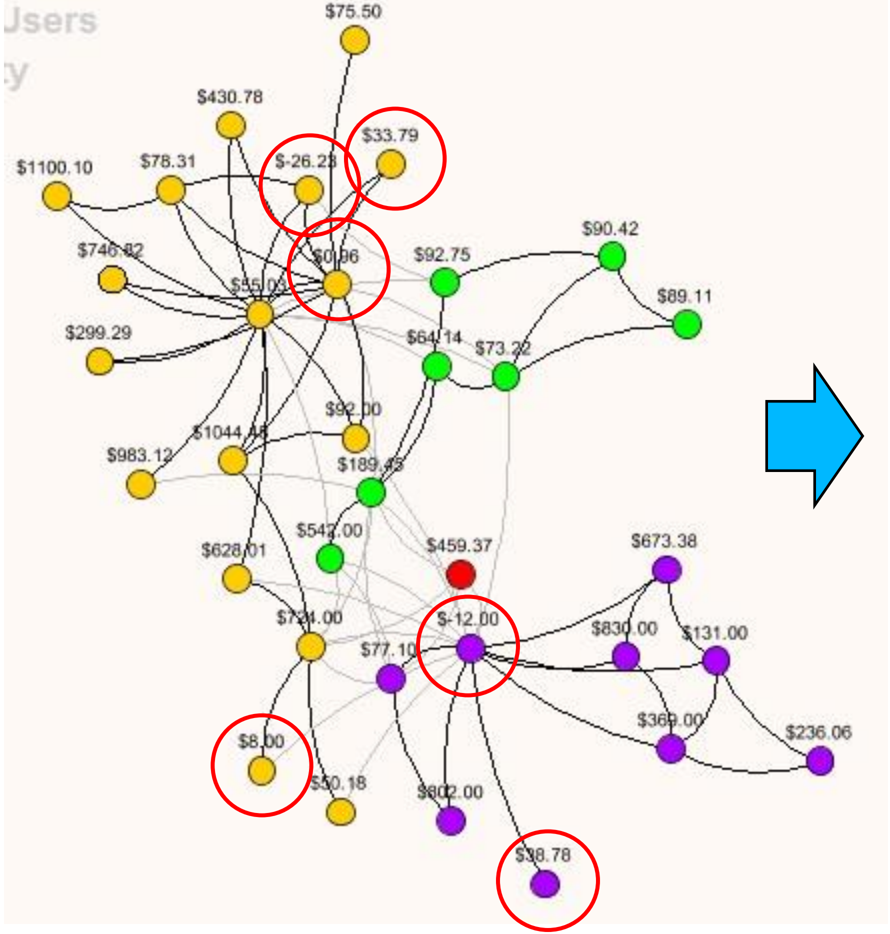
Data Linking and Integration



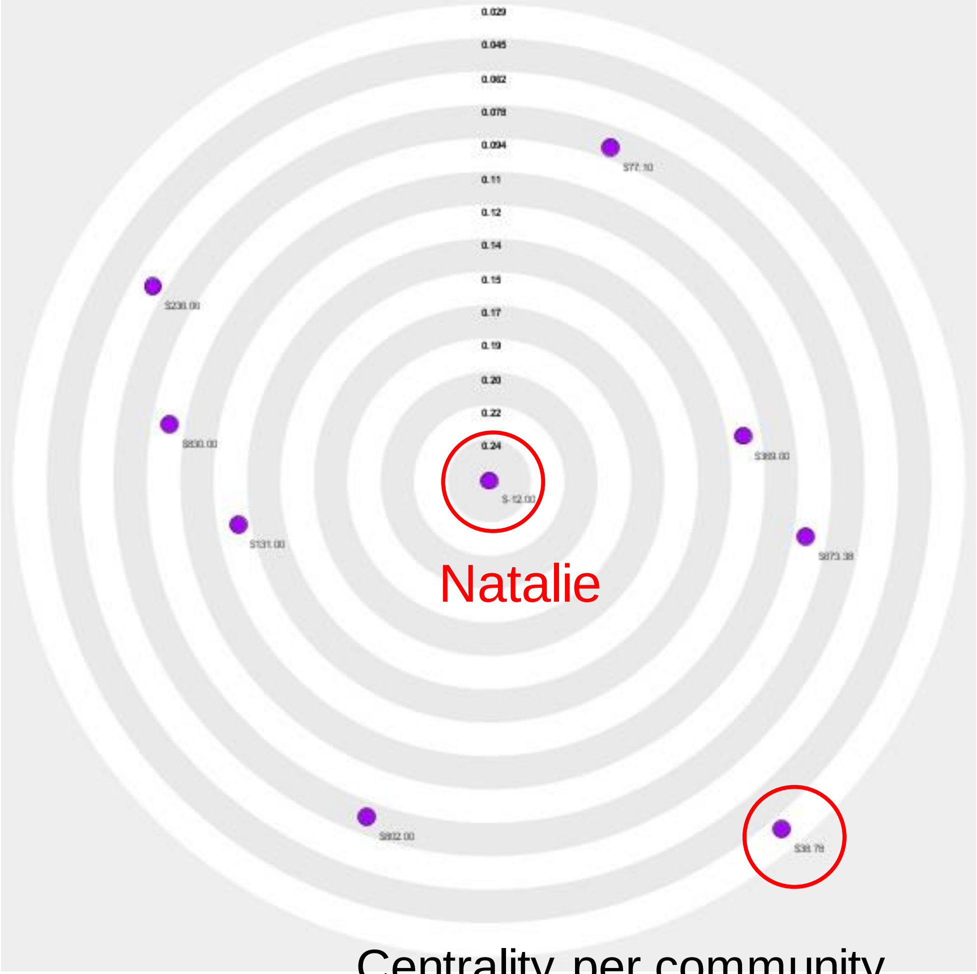
Inter-call network with call frequency



Community Mining

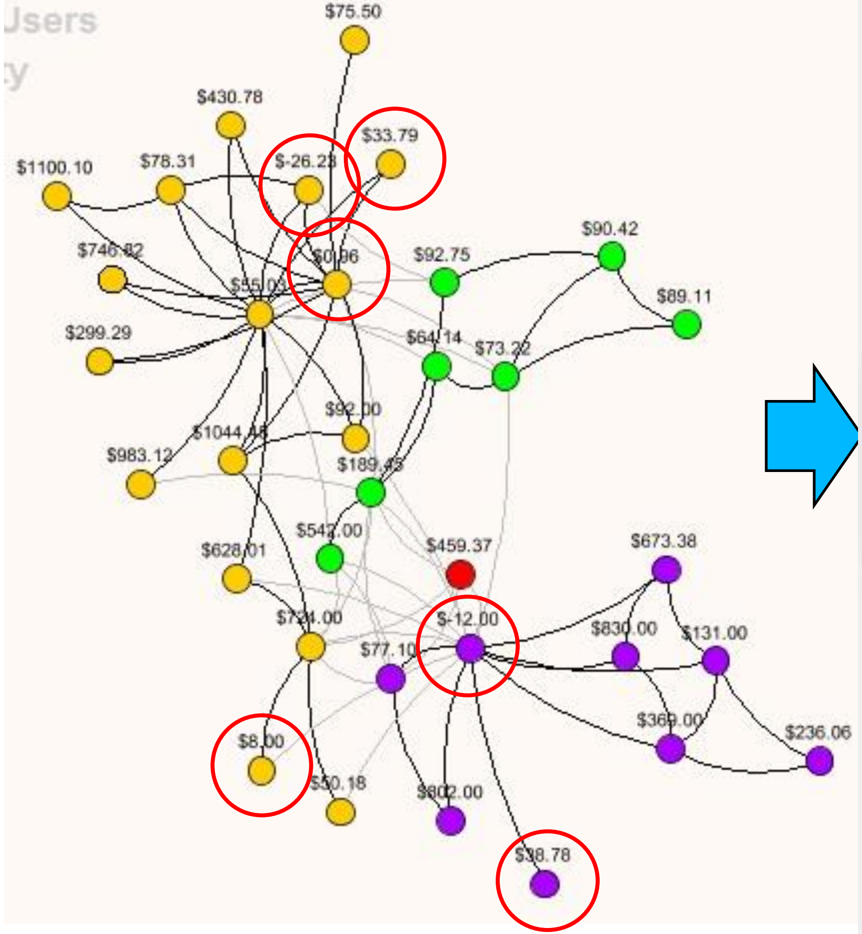


Community Mining

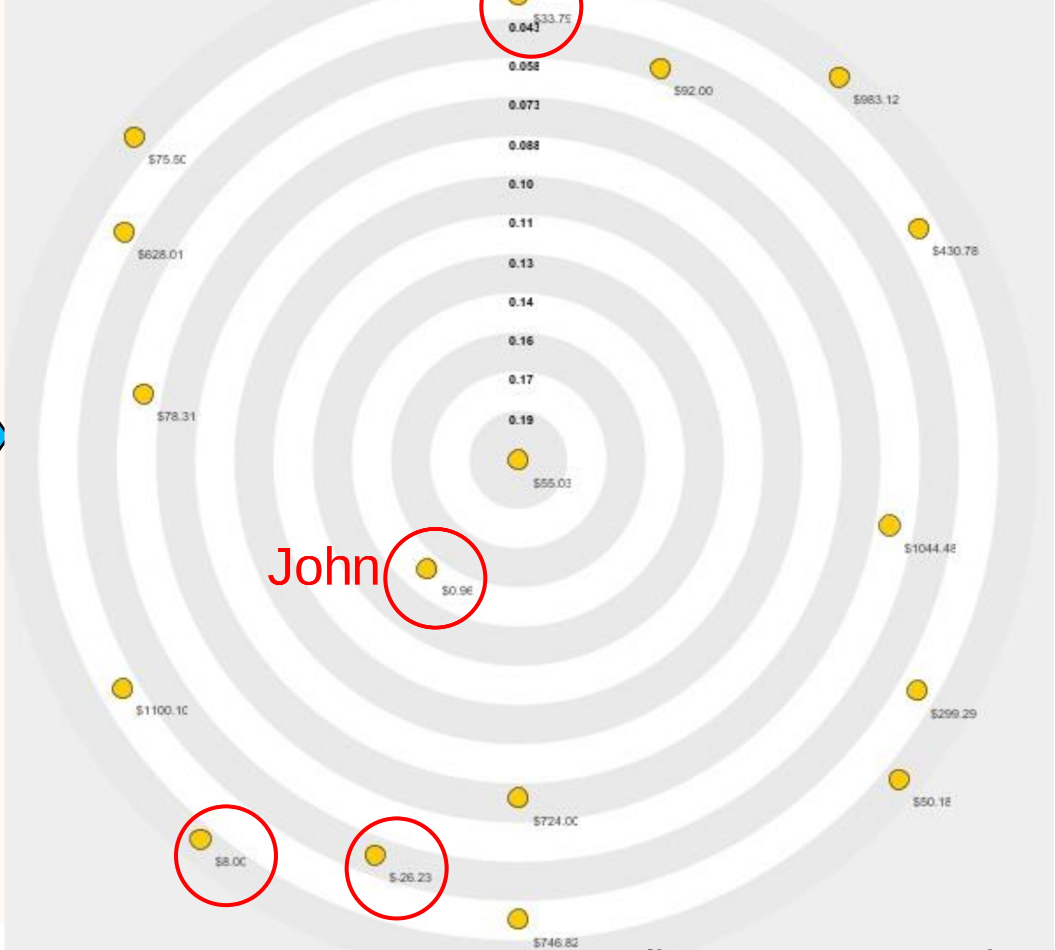


Centrality per community

Dropping Natalie: Risk = $3145.32



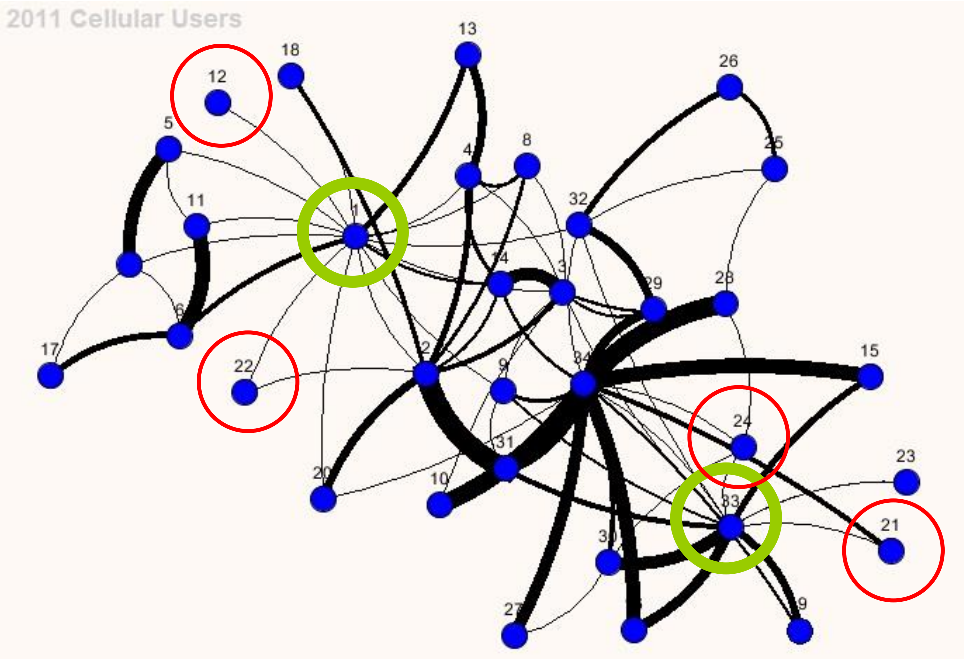
Community Mining



Centrality per community

Dropping John: Risk = $6324.14

ID	Name	Phone Number	City	Plan	Avg. 3m Profit
24	Ben Rikon	403 262 3134	Calgary	3y	($26.23)
1	John Smith	647 225 8085	Toronto	2y	($12)
33	Natalie May	403 409 6223	Calgary	3y	$0.96
22	Wilma Renton	780 118 2388	Edmonton	3y	$8.00
21	Patrick Klum	403 337 9291	Calgary	3y	$33.79
12	Kent Wafegert	647 631 0348	Toronto	3y	$38.78
18	Patty Klien	780 550 1819	Edmonton	1y	$50.18
4	Randy Regal	705 234 6767	Toronto	3y	$77.10
26	Maggie Wong	226 882 0911	Toronto	2y	$89.11
28	Karen Pollonts	403 750 9201	Calgary	3y	$92.75
7	Susan Willcox	780 291 6063	Edmonton	2y	$131.00
3	John Simon	780 886 5053	Edmonton	3y	$189.45
17	Wayne Jones	780 236 3006	Edmonton	3y	$236.06
15	Brent Mavka	403 566 7372	Calgary	2y	$299.29
6	Mary Tasear Smith	780 334 3434	Edmonton	3y	$369.00
8	Martha Witherby	780 322 9768	Edmonton	3y	$459.37
20	Morris Slevchuk	780 434 6280	Edmonton	3y	$628.01
11	Kurt Locke	780 654 1121	Edmonton	3y	$830.00
31	Monica Kwalshuck	403 210 4448	Calgary	3y	$1,044.48

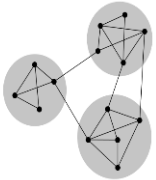


19 customers up for plan renewal
Which one to renew?
Which one to give incentive to stay?

Give incentives to 1 (John Smith -$12) and 33 (Natalie May $0.96) to stay but let the others go.

Exploiting additional data and sophisticated analysis could give a different perspective and provide unexpected insights leading to competitive advantage.

What is a community (cluster in a network)?



Loosely defined as groups of nodes that have relatively more links between themselves than to the rest of the network

- Nodes that have structural similarity (**SCAN**, Xu et al. 2007)
- Nodes that are connected with cliques (**CFinder** by Palla et al. 2005)
- Nodes that a random walk is likely to trap within them (**Walktrap** by Pons and Latapy 2006)
- Nodes that follow the same leader (**TopLeaders**, Rabbany et al. 2010)
- Nodes that make the graph compress efficiently (**Infomap, Infomod**, Rosvall and Bergstrom, 2011)
- Nodes that are separated from the rest by min cut, conductance (flow based methods, e.g. **Kernighan-Lin (KL), betweenness** of Newman)
- Nodes that number of links between them is more than chance (Newman's **Q modularity, FastModularity,** Blondel et al.'s **Louvain)**

Community Mining Algorithms

Different community mining algorithms discover communities from different perspective

How to evaluate and compare the results of different community mining algorithms?

Definition v.s. Evaluation

A congruence relation between defining communities and evaluating community mining results

Q-modularity by Newman and Girvan

- common objective for community detection
- originally proposed to quantify goodness of communities
- still used for evaluating the algorithms

How about Relative Evaluation?

None of the studies on Community Mining Algorithms considers any different validity criteria other than Q-modularity to evaluate the goodness of the detected communities.

Validity criteria defined for clustering evaluation; compares different clusterings of a same data set

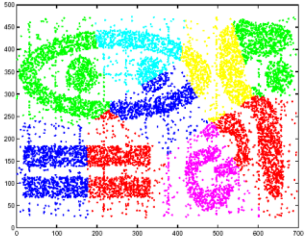


Figure 5: K-means's clustering result on t7.10k.dat with $k = 9$

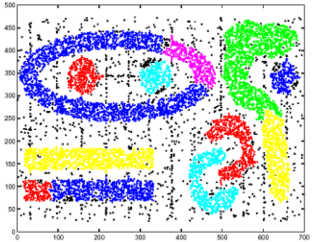


Figure 10: DBSCAN's clustering result on t7.10k.dat with $\epsilon = 5.5$ and MinPts = 4

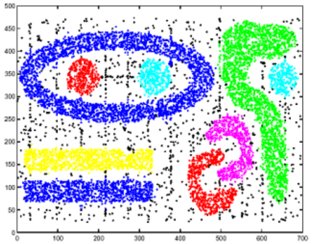


Figure 2: $TURN^*$'s clustering result on t7.10k.dat before cleaning

Clustering quality criteria defined with the assumption that data points consist of vectors of attributes ➔ There is a definition of distance measure (Euclidean or other).

Most clustering quality criteria use averaging between data points to determine a centroid of a cluster

There is no notion Euclidian distance in a graph or the notion of averaged centroid

Internal Evaluation Practice

Generally, an internal criteria quantifies the goodness of a clustering, given only the data (only the *graph* in the case of communities).

- ➢ makes assumption about what are good communities ⇒ is not appropriate to validate results of algorithms built upon different assumptions (e.g. are not optimizing Q)
- ➢ Not a fair eval

Internal Evaluation Practice (Cont.)

Different objectives for internal/relative evaluation (Q, VRC, Silhouette, etc.) perform differently in different settings ⇒ No overall winner.

An internal evaluation criterion encompasses the same non-triviality as of the community mining task itself

External Evaluation

Validating on a set of benchmarks with known ground-truth communities.

➢ Few and typically small real world benchmarks

⇒ **Synthetic** benchmarks or on large real networks with **explicit or predefined communities**

Synthetic Benchmarks

Performance of an algorithm on synthetic benchmarks is a predictor of its performance on real networks

Only true if synthetic benchmarks are realistic

- ➢ The current common generators, e.g. LFR, are far from characteristics of the real networks

Generating Attributed Networks with Communities, PLoS One. 2015 Apr 20;10(3)

Attributes as Benchmark

Alternative to synthetic benchmarks?

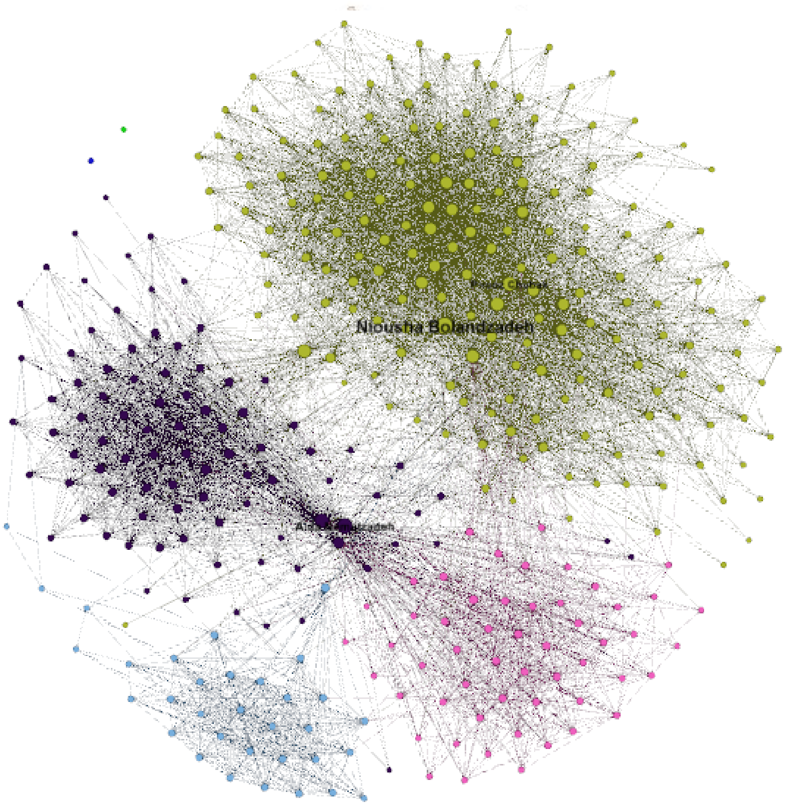
Large real networks with ground-truth defined based on explicit properties of nodes (e.g. SNAP)

- **venues** in collaboration network of authors from DBLP,
- product **categories** in Amazon co-purchasing network

This ground-truth is imperfect and incomplete [Cunnigham 2013]

⇒ metadata or labeled attributes **correlated** with the underlying communities

Correlation of Communities and Attributes



User attributes can act as the primary organizing principle of the communities

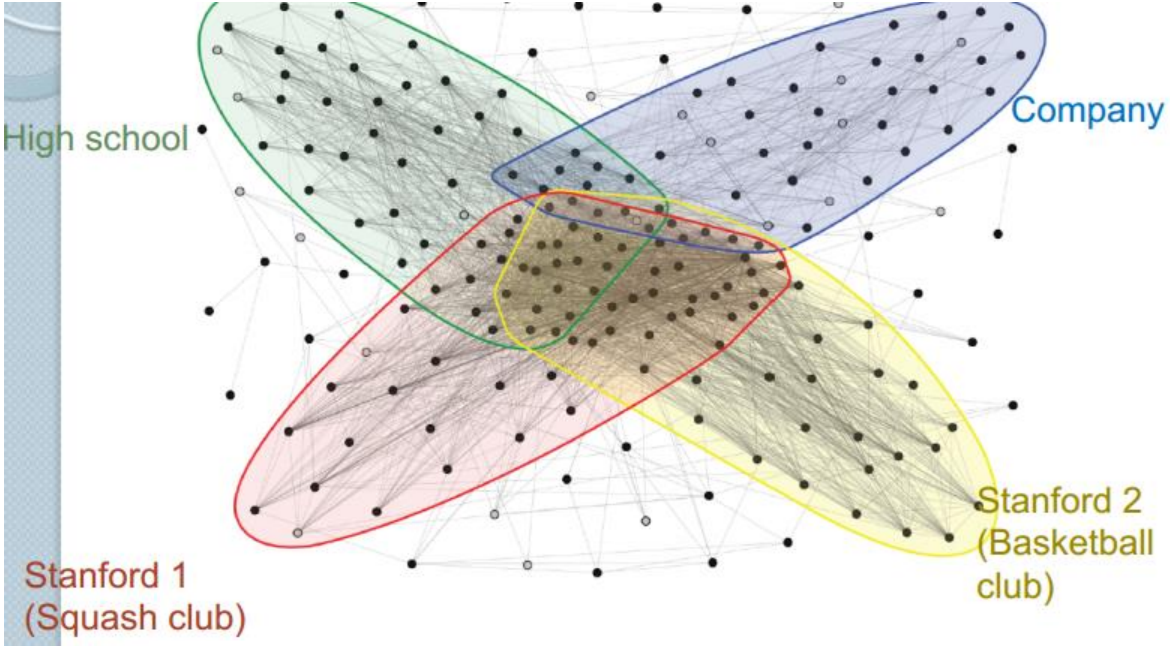
Amanda L Traud, Eric D Kelsic, Peter J Mucha, and Mason A Porter. **Comparing community structure to characteristics in online collegiate social networks.** SIAM review, 53(3): 526–543, 2011.

Correlation significantly depends on this agreement index and differs significantly even between those indices have been known to be linear transformation of each other

Figs from Guo et.al. 2011

Correlation of Communities and Attributes

Jaewon Yang and Jure Leskovec. **Defining and evaluating network communities based on ground-truth.** In Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics, page 3. ACM, 2012



imperfect and incomplete (Lee and Cunningham (2013))

Figs from Guo et.al. 2011

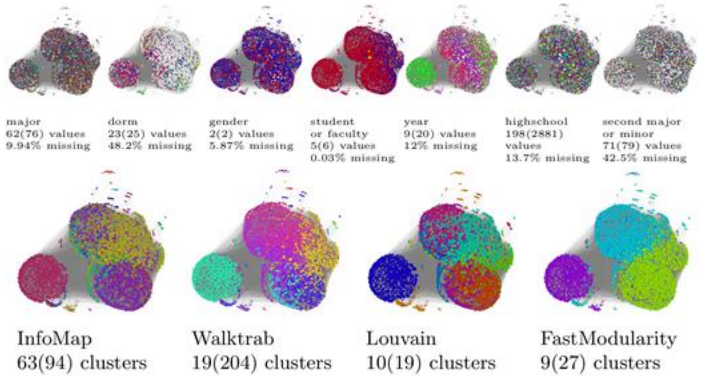
Study

- Investigates correlations between attributes and community structure
  - Using our network specific clustering agreement indexes
- Presents **community guidance by attributes**
  - We guide our TopLeaders community detection method to *find the right number of communities based on the available attributes data*

Correlation of Communities and Attributes

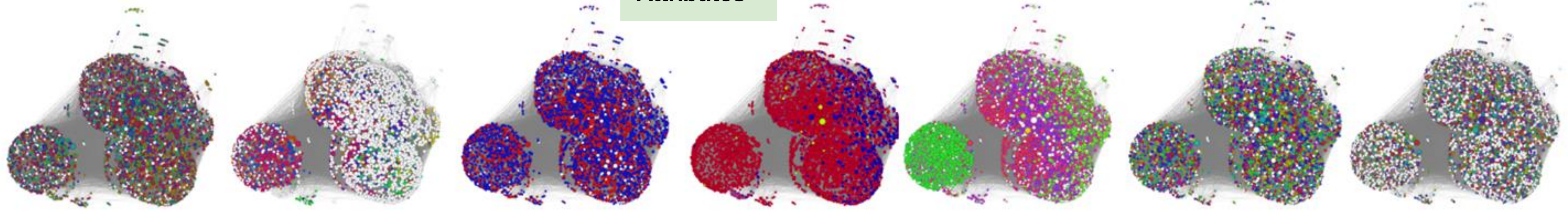
Facebook friendship network

- for 100 US universities
- each node has 7 attributes



We compare correlation of the results from four different community mining algorithms, with each attribute in the dataset (InfoMap, WalkTrap, Louvain, FastModularity)

Attributes



major
62(76) values
9.94% missing

dorm
23(25) values
48.2% missing

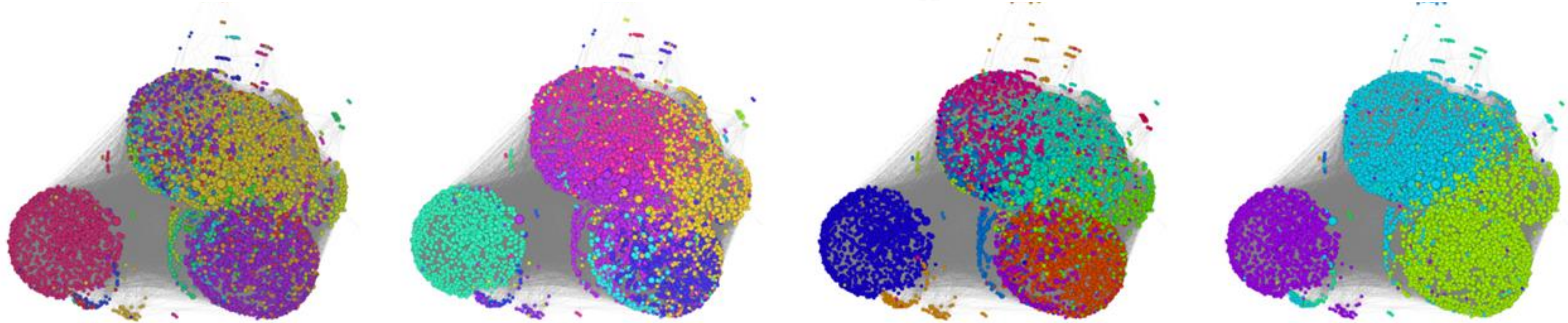
gender
2(2) values
5.87% missing

student
or faculty
5(6) values
0.03% missing

year
9(20) values
12% missing

highschool
198(2881)
values
13.7% missing

second major
or minor
71(79) values
42.5% missing



InfoMap
63(94) clusters

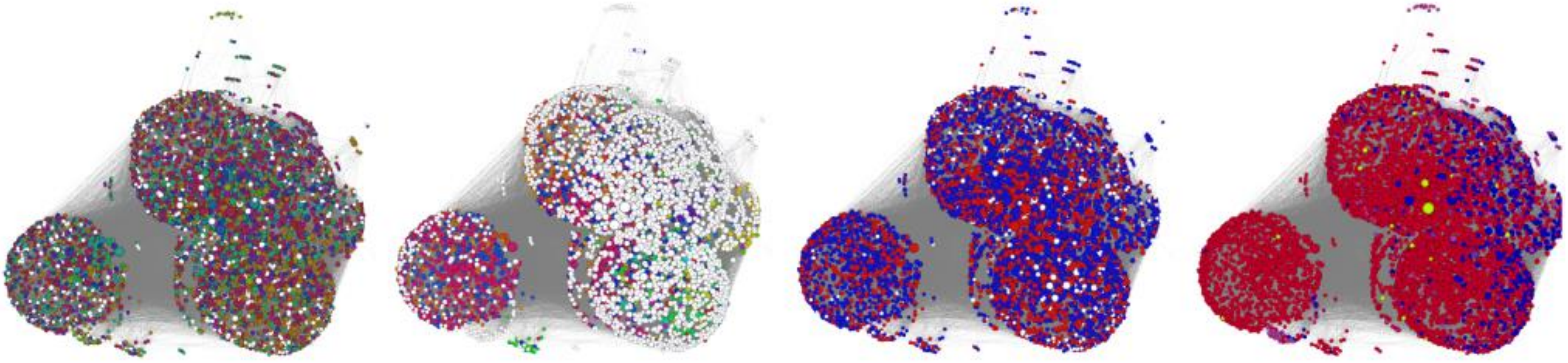
Walktrab
19(204) clusters

Louvain
10(19) clusters

FastModularity
9(27) clusters

Communities

Zoomed



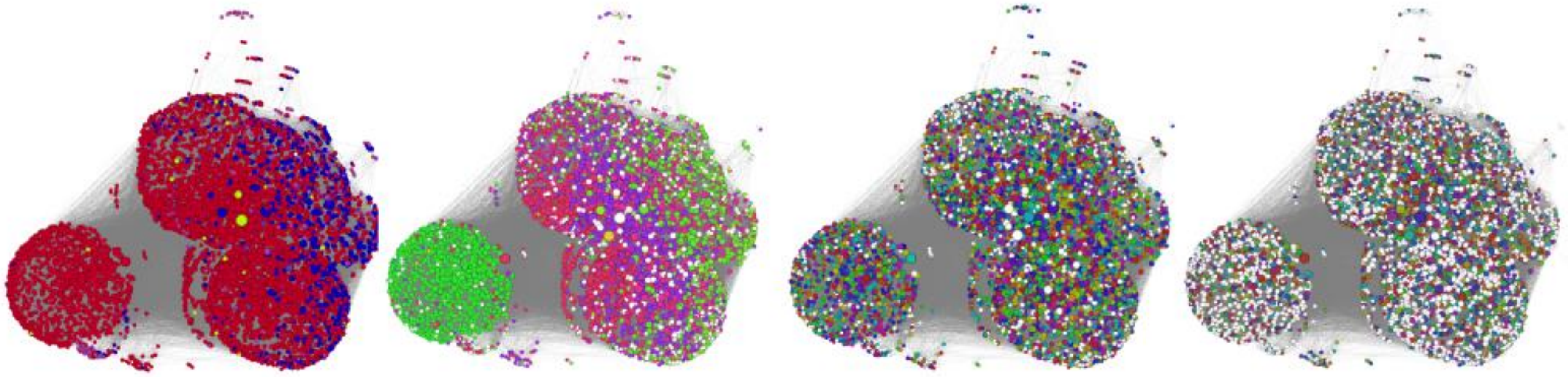
major
62(76) values
9.94% missing

dorm
23(25) values
48.2% missing

gender
2(2) values
5.87% missing

student
or faculty
5(6) values
0.03% missing

Zoomed



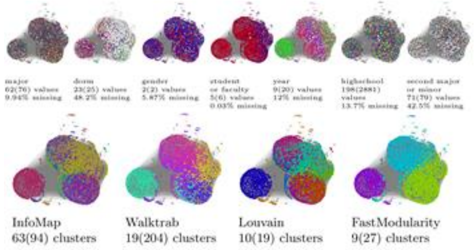
student
or faculty
5(6) values
0.03% missing

year
9(20) values
12% missing

highschool
198(2881)
values
13.7% missing

second major
or minor
71(79) values
42.5% missing

Correlation of Communities and Attributes



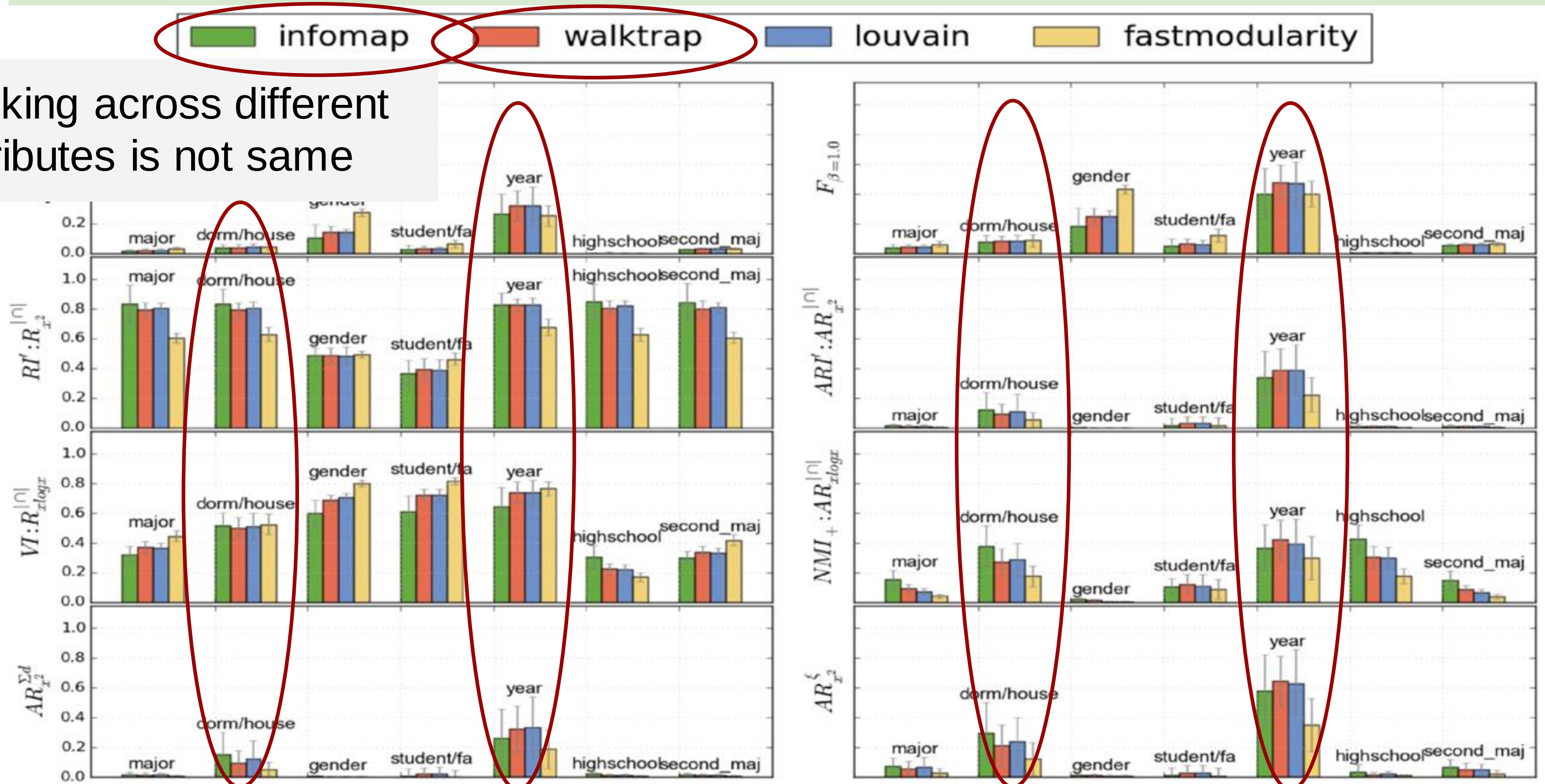
The correlation are measured using clustering agreement indices

- Unique attribute values ⇒ clustering
- Eight agreement indices
  - Jaccard Index, F-measure, Variation of Information(VI), Normalized Mutual Information(NMI), Rand Index(RI), Adjusted Rank Index(ARI),
  - Two structure based extensions of ARI tailored for comparing network clusters with overlap function as
    - the sum of weighted degrees
    - the number of common edges

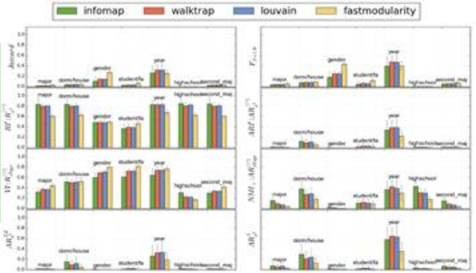
"Generalization of Clustering Agreements and Distances for Overlapping Clusters and Network Communities." *arXiv preprint arXiv:1412.2601* (2014).

Ranking of Algorithms averaged over all Facebook 100 dataset

ranking across different attributes is not same



Ranking of Algorithms



Attributes and communities **are correlated**

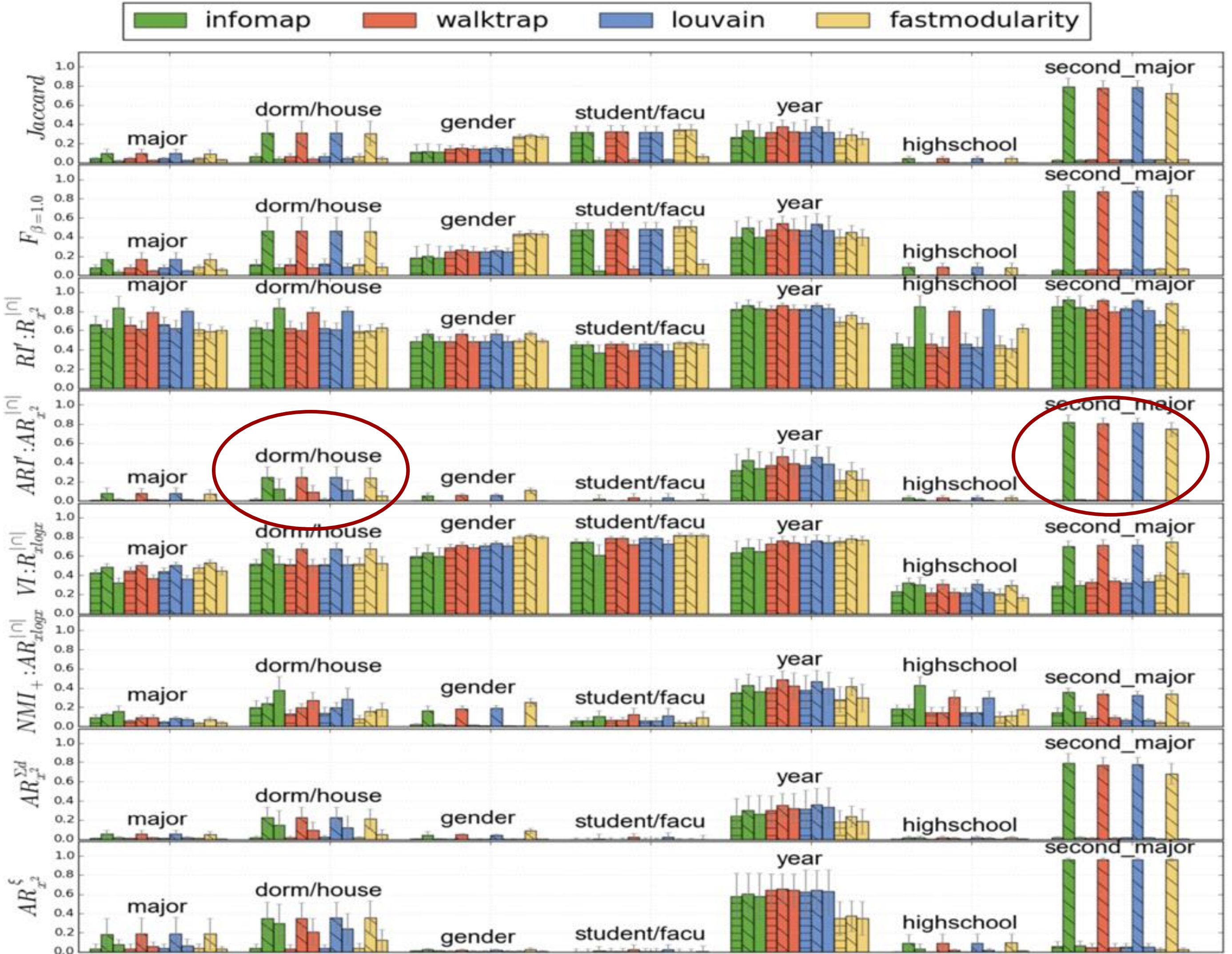
But it is not wise to compare the general performance of community mining algorithms based on their agreements with a selected attribute as the ground-truth

- Instead one should treat attributes as another source of information
  - to fine tune the parameters of a community mining algorithm, so that it results in a community structure which compiles most with our selected attribute

Missing Values

→ horizontal: removing missing values
→ diagonal: adding missing values as a single cluster
→ solid: lifting the covering assumption (our formulation)

Significant difference in agreements based on how we treat missing values



Influence & Selection

*The **relations** between nodes **motivates** them to **develop similar attributes (influence)**, a property known as social influence, whereas the **similarities** between them **motivates** them to **form relations (selection)**, a property referred to as homophily.*

Also explains the correlations observed

In Presence of Attributes

Groupings that are both internally well connected and having homogeneous attributes

- structural attribute clustering [Zhou et al. 2009]
- cohesive patterns mining [Moser et al. 2009]

⇒ Combining attribute and link data, rather than validating one based on the other

Community guidance by attributes:

attribute is used to direct a community mining algorithm

Community Guidance by Attributes

- Guide TopLeaders to find the right number of communities, based on the agreements of its result with the given attribute
  - The number of communities, k for short, is the main parameter for the TopLeaders algorithm, similar to the k-means algorithm for data clustering

Top Leaders Community Detection Approach in Information Networks, SIGKDD SNA-KDD Workshop 2010

- The concept is however general and can be applied to fine tune the parameters of any community mining algorithm

Top Leaders Approach

Top Leaders Community Detection Approach in Information Networks, SIGKDD SNA-KDD Workshop 2010

A leader is the most central member in a community

Algorithm 1 Top Leaders algorithm

Input: A social network G, and k the number of desired communities

initialize k leaders
repeat
  {finding communities}
  for all Node n $\in$ G **do**
    if n $\notin$ leaders **then**
      associate n to a leader {Algorithm 2}
    end if
  end for
  {updating leaders}
  for all l $\in$ leaders **do**
    l $\leftarrow \arg\max_{n \in Community(l)} Centrality(n)$
  end for
until there is no change in the leaders

Associating Nodes to Leaders

Algorithm 2 Associate n to its leader

Input: Social network G, node n, set of k leaders

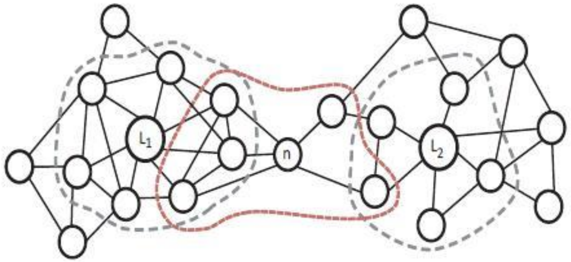
depth $\leftarrow$ 1
CanList $\leftarrow$ leaders
repeat

$$\text{CanList} \leftarrow \underset{\substack{c \in CandList \,\wedge \\ |\aleph(n_1,d) \cap \aleph(n_2,d)| > \gamma}}{\arg\max} |\aleph(n_1, d) \cap \aleph(n_2, d)|$$

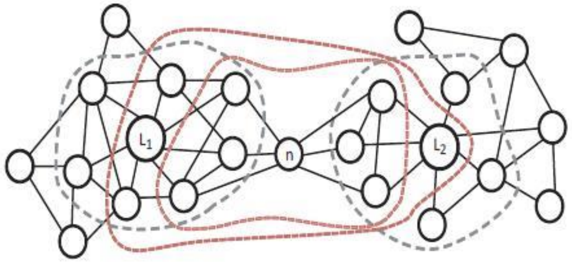
depth $\leftarrow$ depth+1
until $|\text{CanList}| \leq 1 \vee \text{depth} > \delta$

if $|\text{CanList}| = 0$ **then** {No candidate leader}
associate n as an outlier
else if $|\text{CanList}| > 1$ **then** {Many candidates}
associate n as a hub
else {Only one candidate leader in CanList}
associate n to CanList
end if

Community membership of the nodes is association of followers to nearby leaders

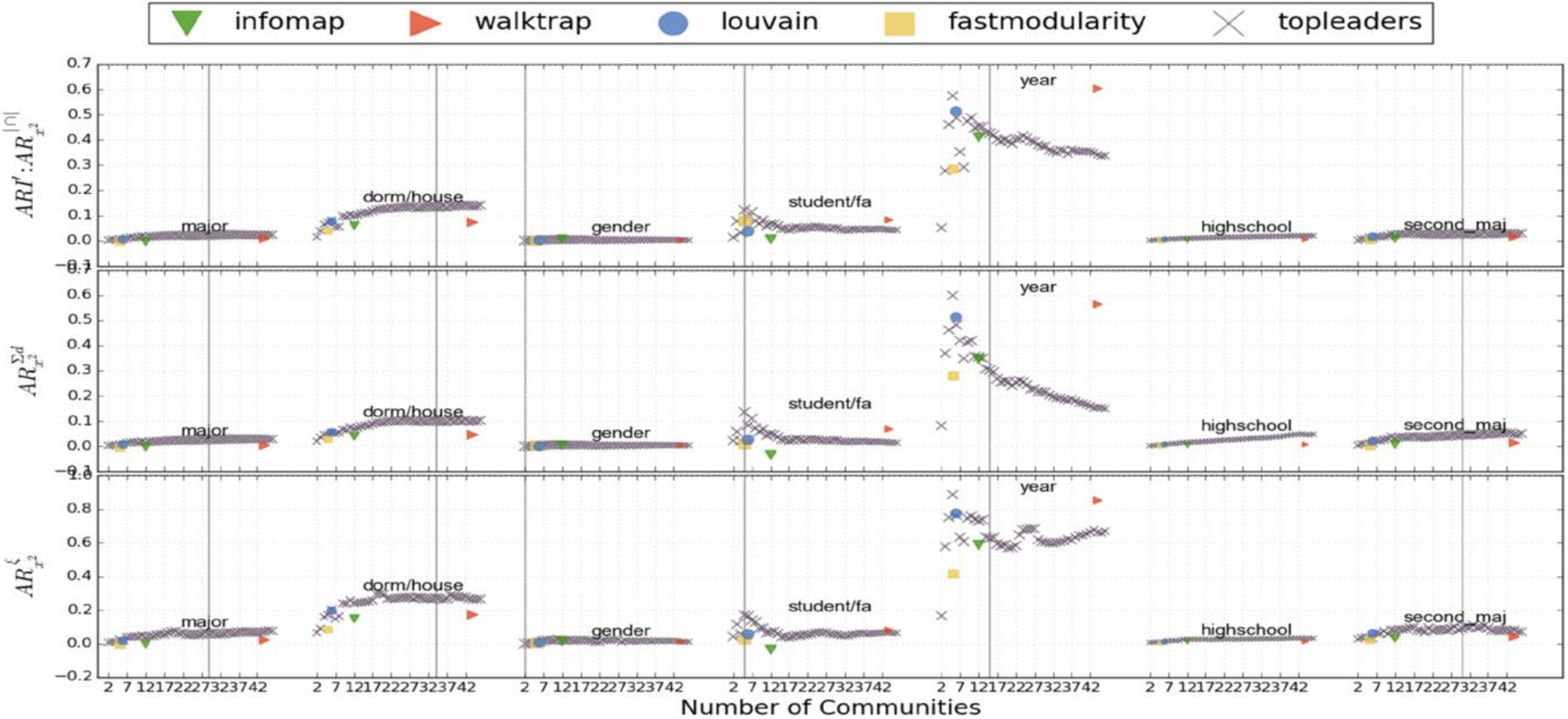


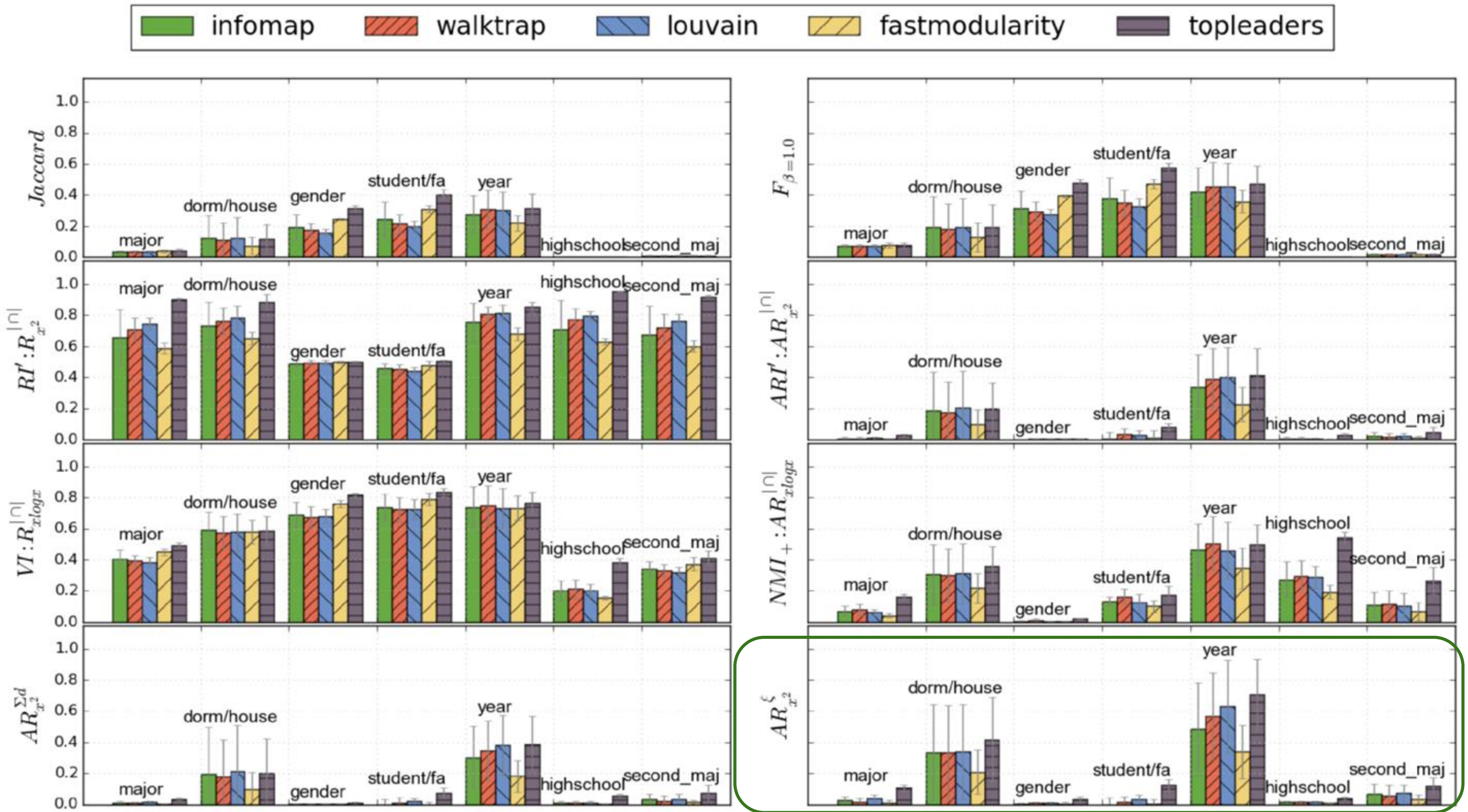
(a) Intersection of neighbourhoods



(b) Expanding Neighbourhoods

Finding k, the number of clusters





infomap
walktrap
louvain
fastmodularity
topleaders
Jaccard
$F_{\beta=1.0}$
$RI':R_{x^2}^{|\cap|}$
$ARI':AR_{x^2}^{|\cap|}$
$VI:R_{xlogx}^{|\cap|}$
$NMI_+:AR_{xlogx}^{|\cap|}$
$AR_{x^2}^{\Sigma d}$
$AR_{x^2}^{\xi}$
major
dorm/house
gender
student/fa
year
highschool
second_maj

Conclusions & Future Works

- Different evaluation approaches for community detection
- Correlation between characteristics of nodes and their connections
- Proposed the concept of **community guidance by attributes**
  - algorithm guided to communities corresponding most to a given attribute
  - useful in real world, since we often have access to both link and attribute information, and an idea of how communities will be used
    - For example, communities in PPI networks are correlated with functional categories of their members, which are used to predict the previously uncharacterized protein complexes; in such case, one might be interested to select the community structure that corresponds most with the available functional categories