Attributed Networks:
Social Circles, Summarization, Comparison

Leman Akoglu

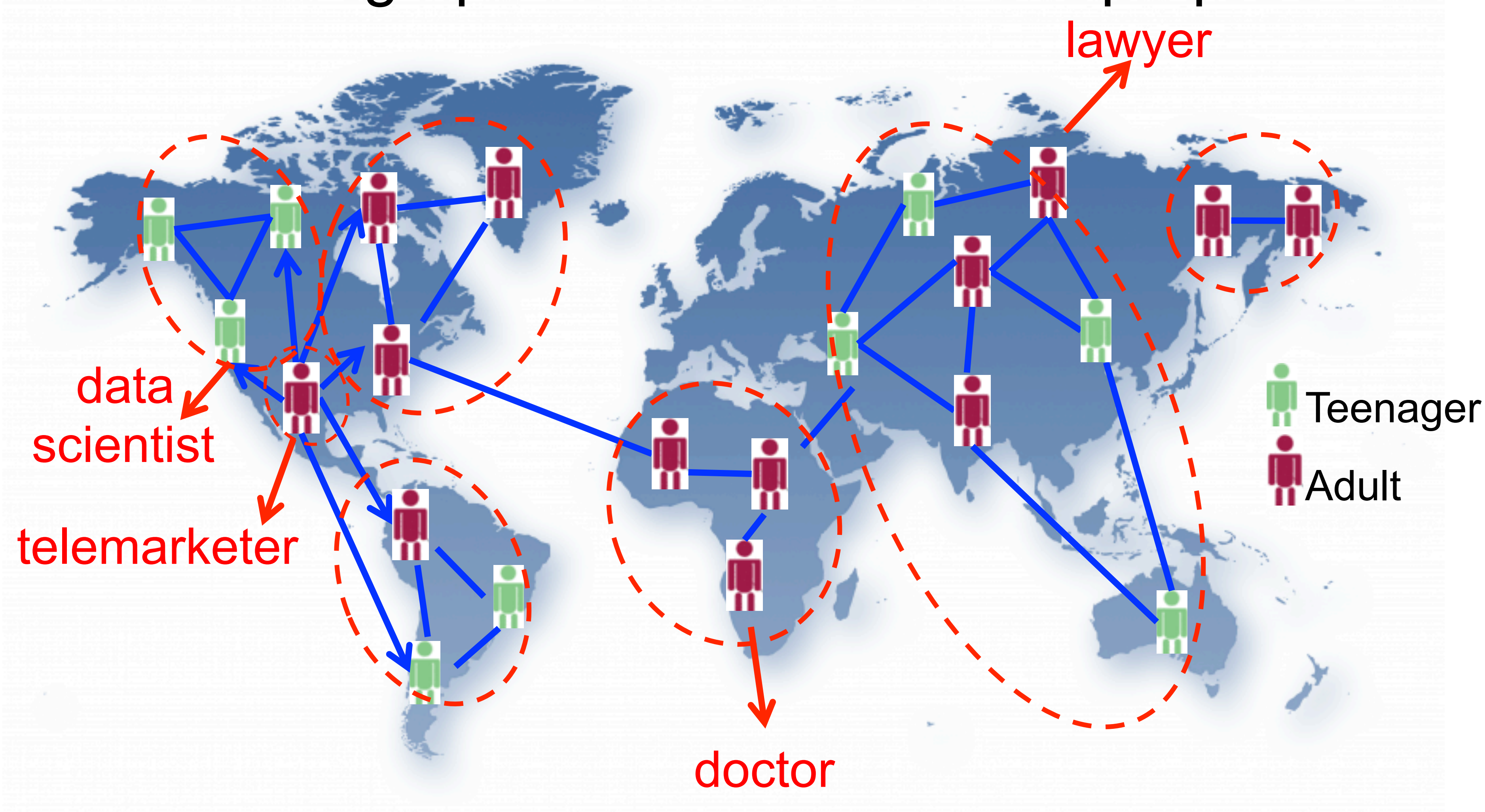
Joint work with Bryan Perozzi (Google Research NYC),
Rashmi Raghunandan, Shruti Sridhar, Upasna Suman (CMU),
Aria Rezaei (Stony Brook University).

NetSci 2018 Satellite on
Machine Learning In Network Science
June 12, 2018

Carnegie Mellon University
Heinz College

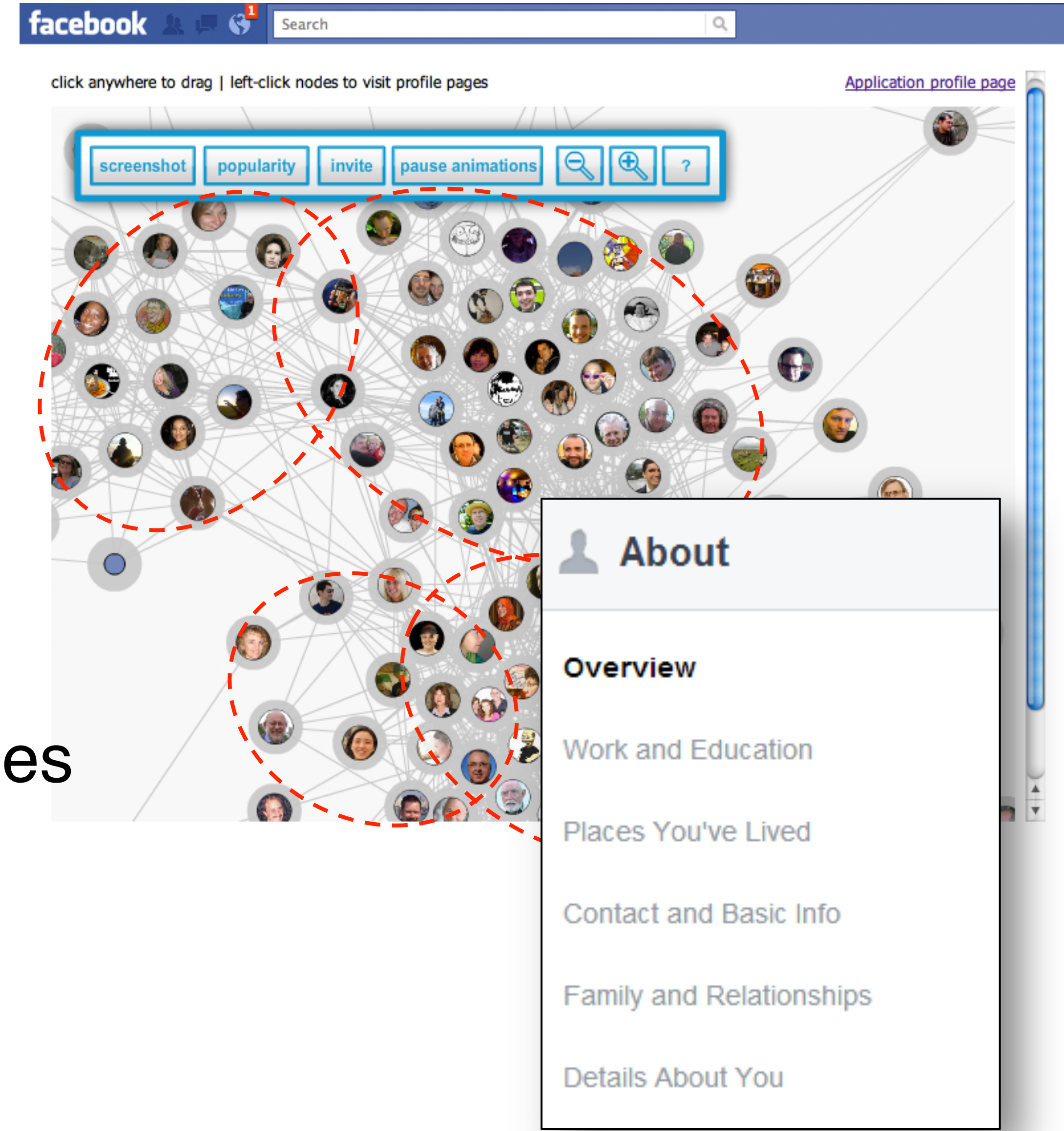
Attributed graphs

Attributed graph: each node has 1+ properties



Attributed networks

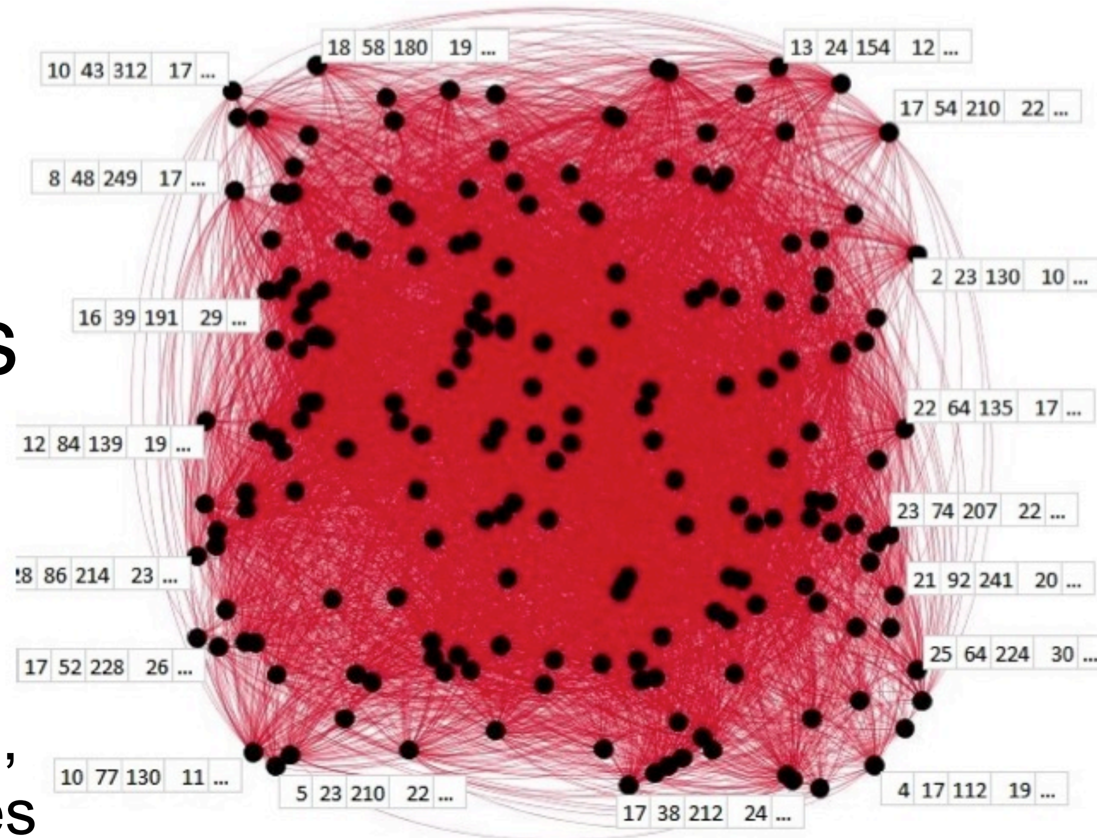
- Social networks
  - demographics, lifestyles, likes, …
- PPI networks
  - Gene encodings
- Gene interaction networks
  - ontological properties
- Web
  - page properties
- …



Motivating question:

How can we **make sense** of node-attributed networks **?**

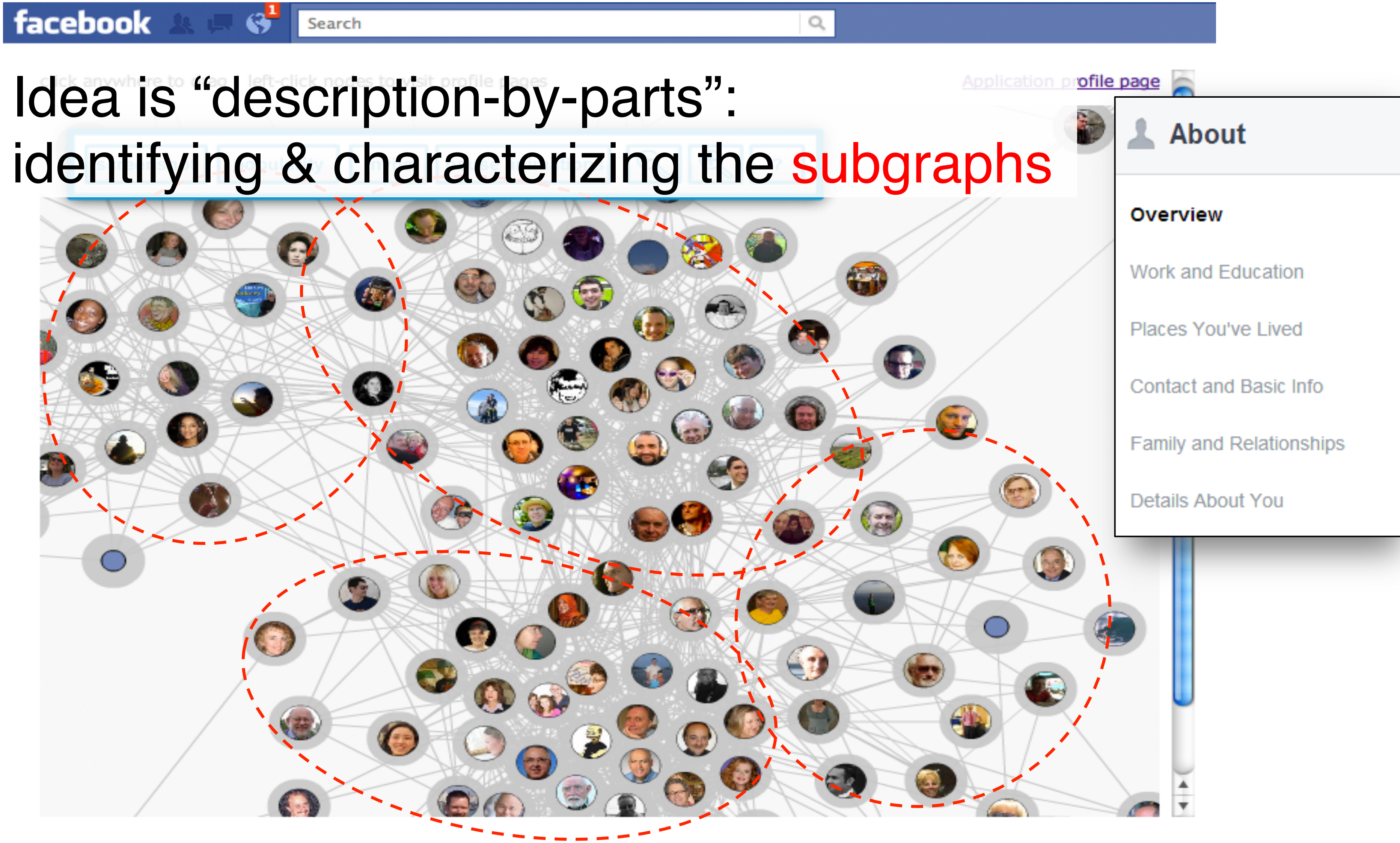
- subgraphs
- summaries
- comparisons

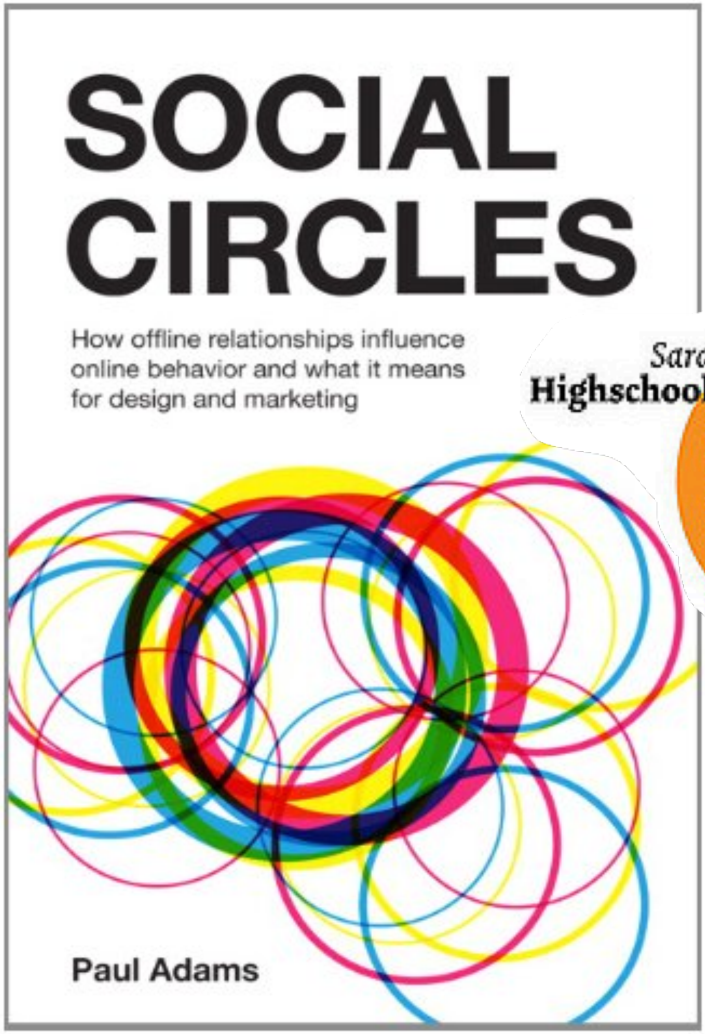


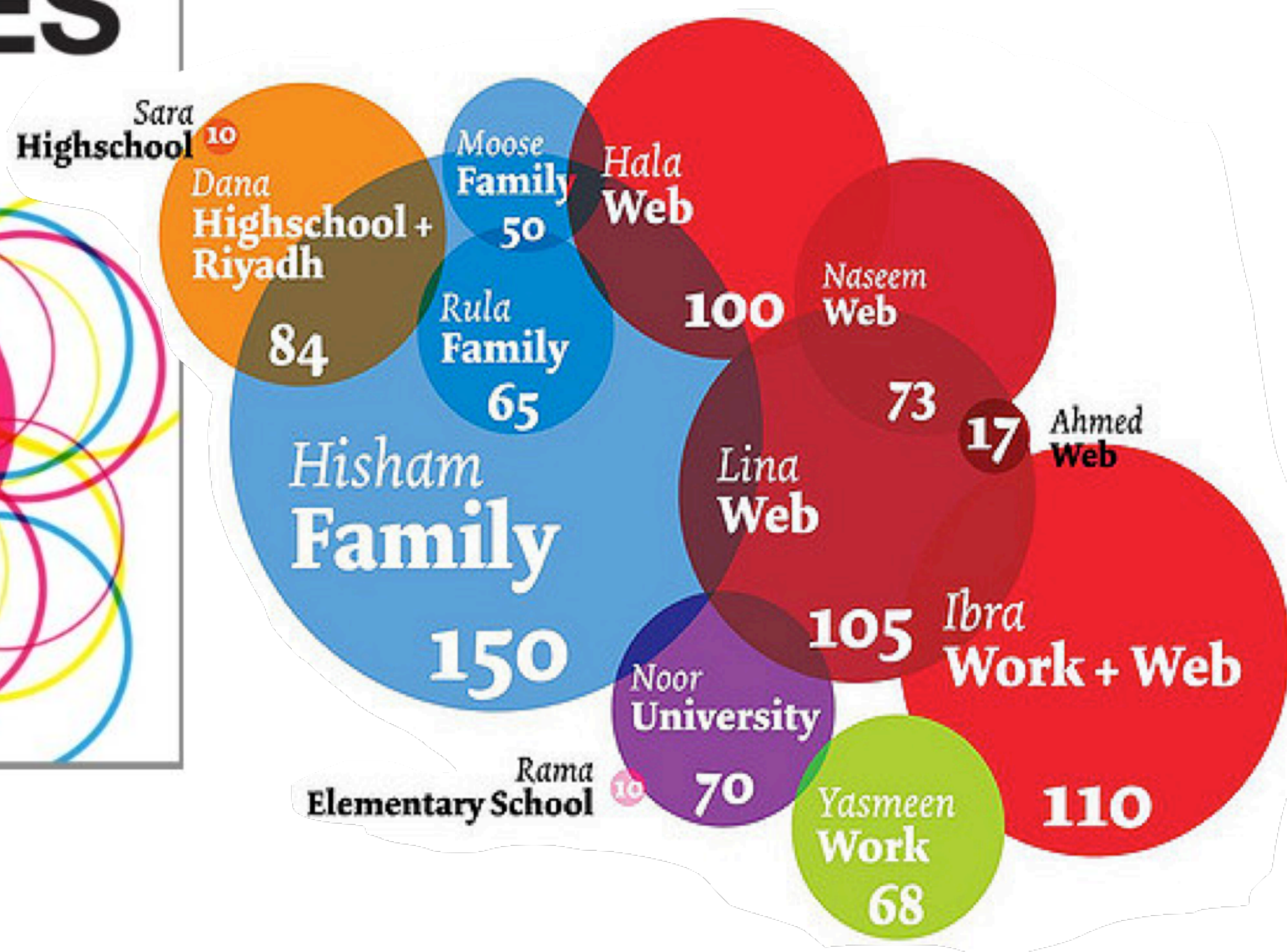
220 nodes,
6215 edges

Attributed networks

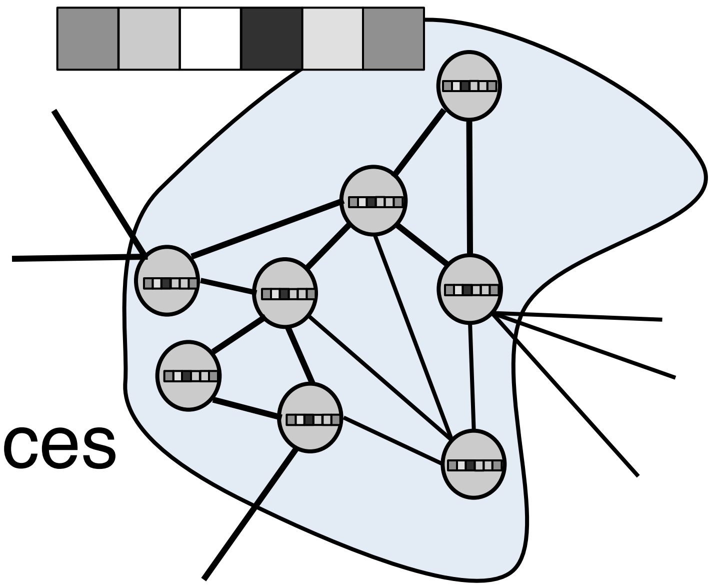
Idea is "description-by-parts":
identifying & characterizing the subgraphs




SOCIAL CIRCLES
How offline relationships influence online behavior and what it means for design and marketing
Paul Adams


Sara
Highschool
10
Dana
Highschool + Riyadh
84
Moose
Family
50
Rula
Family
65
Hisham
Family
150
Hala
Web
100
Naseem
Web
73
17
Ahmed
Web
Lina
Web
105
Ibra
Work + Web
110
Noor
University
70
Rama
Elementary School
10
Yasmeen
Work
68

Research questions:



① How to characterize & **measure the quality** of …

② How to **summarize** & **interactively explore** …

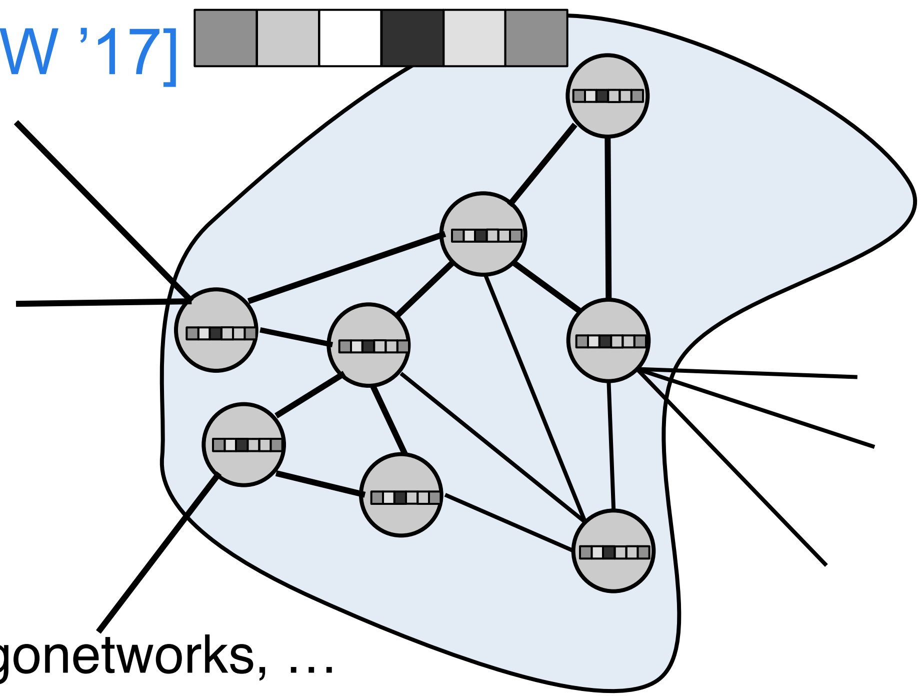
③ How to **characterize** differences between **classes** of …

… attributed subgraphs?

1) **Scalable Anomaly Ranking of Attributed Neighborhoods** SIAM SDM 2016
2) **Discovering Communities and Anomalies in Attributed Graphs: Interactive Visual Exploration and Summarization** ACM TKDD, 2018
   Bryan Perozzi and Leman Akoglu
3) **Ties That Bind - Characterizing Classes by Attributes and Social Ties**
   Aria Rezaei, Bryan Perozzi, Leman Akoglu WWW 2017 Companion

This talk

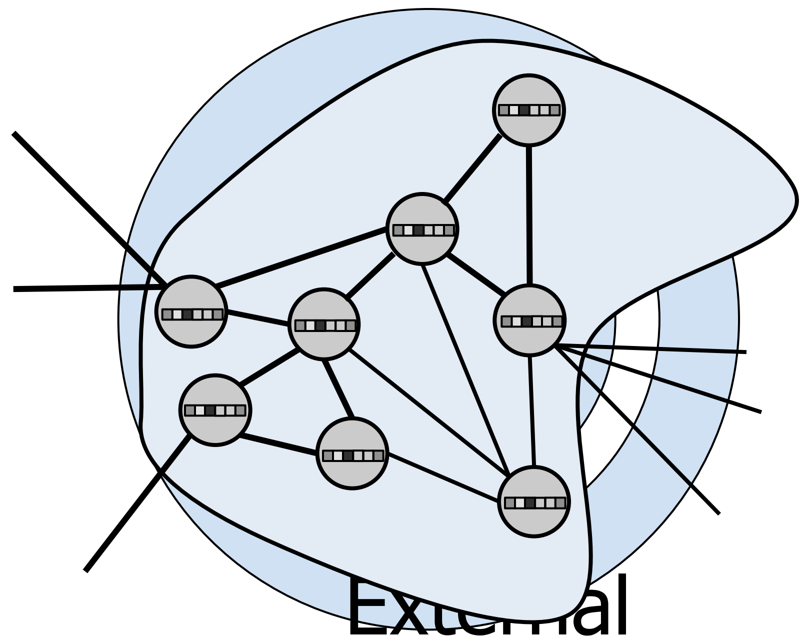
- Attributed (sub)graphs*
  - Subgraphs [SIAM SDM'16]
  - Summarization [ACM TKDD'18]
  - Comparisons [WWW '17]



* social circles, communities, egonetworks, …

What's a "good" subgraph anyway?

- **Given** an attributed subgraph, how to **quantify** its quality?
  - Structure-only
    - Internal-only
      - average degree
    - Boundary-only
      - cut edges
    - Internal + Boundary
      - conductance
  - Structure + Attributes

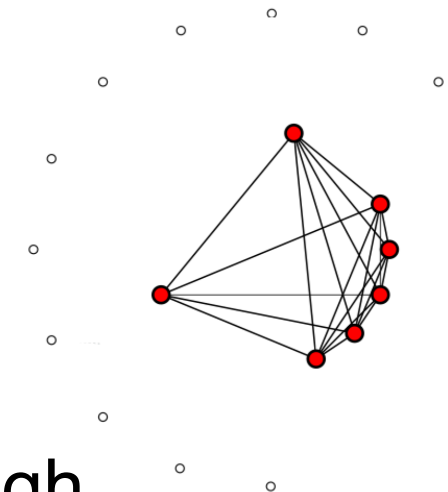


Scalable Anomaly Ranking of Attributed Neighborhoods

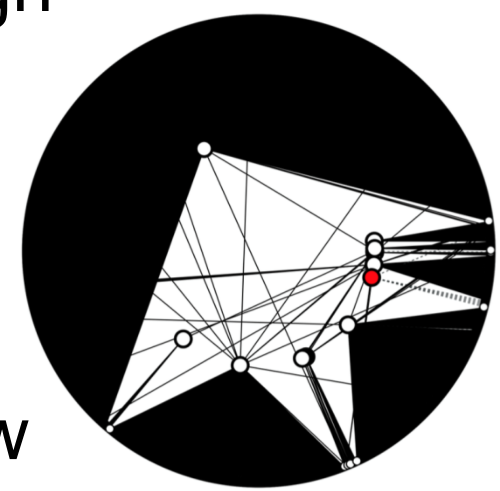
Bryan Perozzi and Leman Akoglu SIAM SDM 2016.

Normality (intuition)

- **Given** an attributed subgraph how to **quantify** quality?
  - Internal
    - structural density



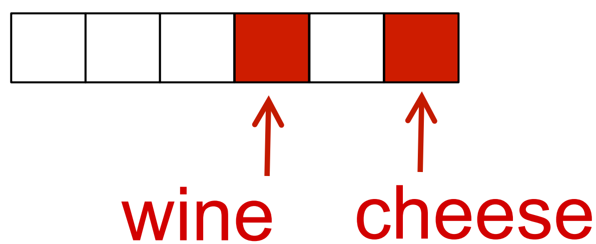
high

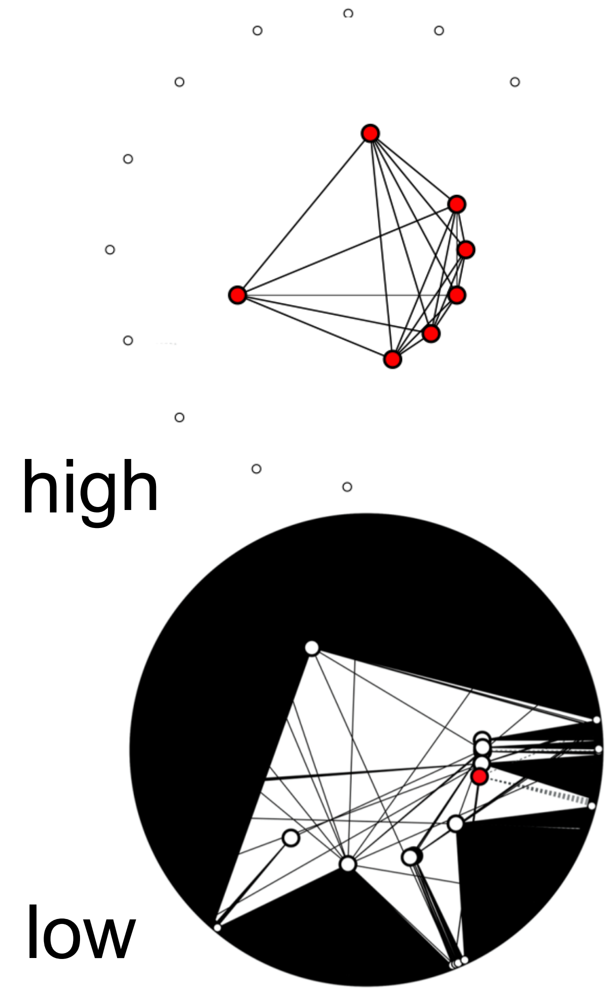


low

Normality (intuition)

- **Given** an attributed subgraph how to **quantify** quality?
  - Internal
    - structural density AND
    - attribute coherence
      - *neighborhood **"focus"***

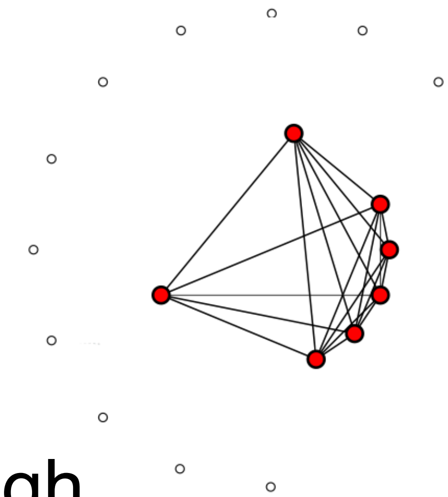




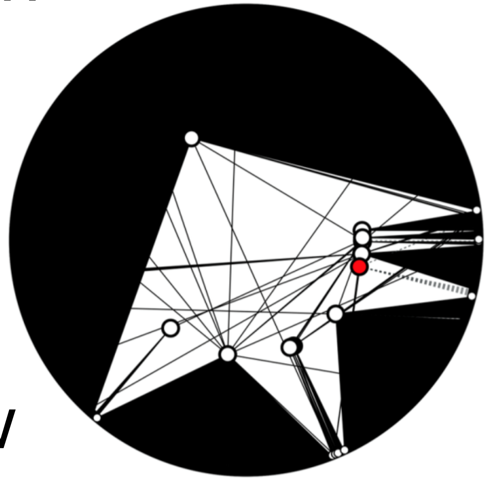
Normality (intuition)

- **Given** an attributed subgraph how to **quantify** quality?
  - Internal
    - structural density AND
    - attribute coherence
      - *neighborhood **"focus"***
  - Boundary
    - structural sparsity, OR
    - external separation
      - ***"exoneration"***

high

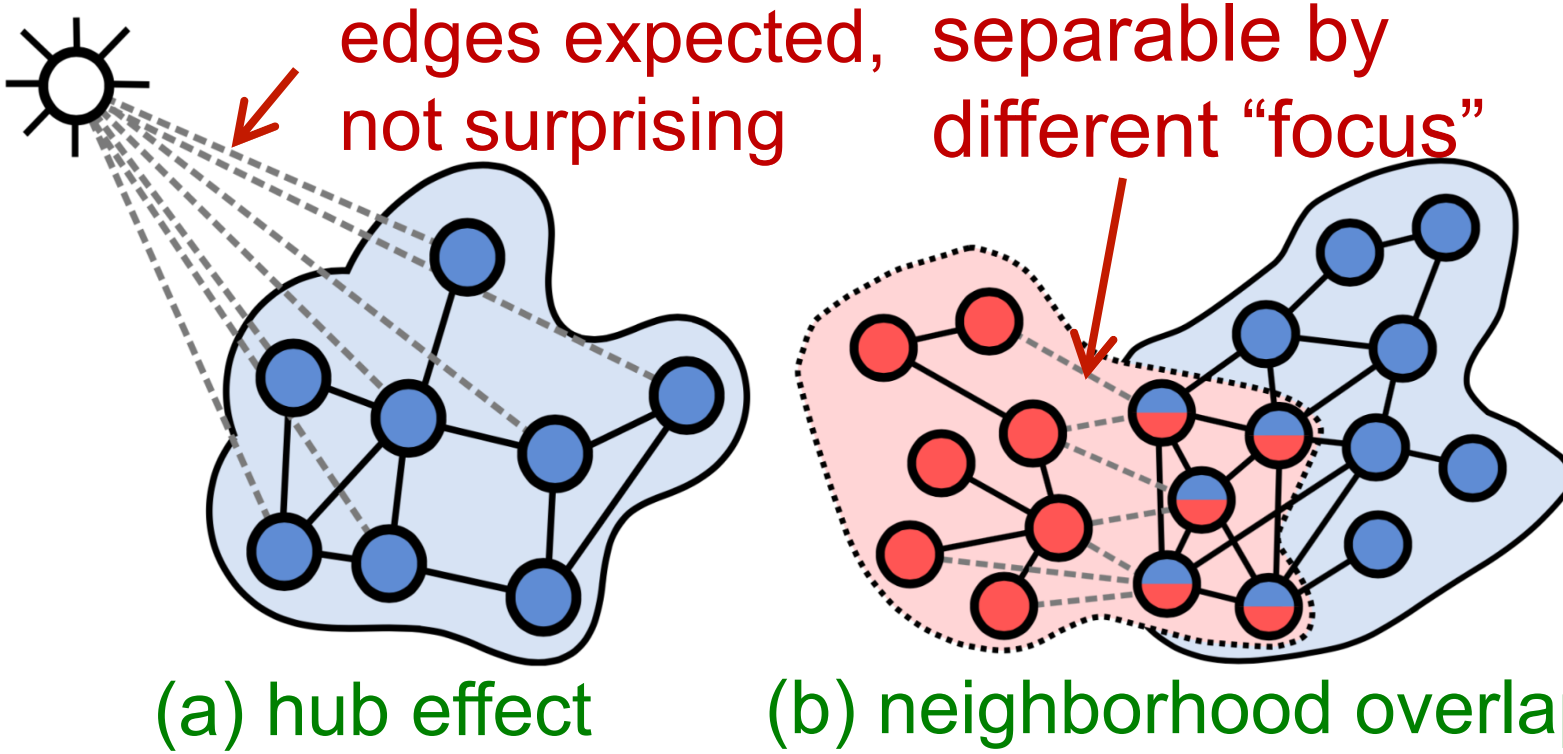


low



Normality (intuition)

- ***"exoneration"*** : by (a) null model, (b) attributes



- Motivation:
  - no good cuts in real-world graphs [Leskovec+ '08]
  - social circles overlap [McAuley+ '14]

The measure of Normality

Null model

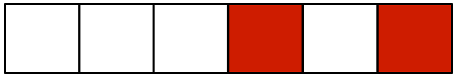
$$\underline{N} = \boxed{I} + E = \boxed{\sum_{i \in C, j \in C} \left(A_{ij} - \frac{k_i k_j}{2m}\right) s(\mathbf{x_i}, \mathbf{x_j} | \mathbf{w})}$$

$$- \sum_{\substack{i \in C, b \in B \\ (i,b) \in \mathcal{E}}} \left(1 - \min\left(1, \frac{k_i k_b}{2m}\right)\right) s(\mathbf{x_i}, \mathbf{x_b} | \mathbf{w})$$

internal consistency

dot-product, or Kronecker's δ

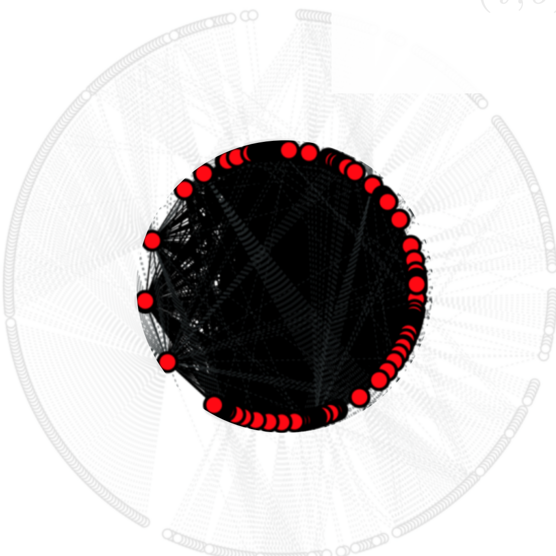
"focus" vector



wine cheese

1

Details



The measure of Normality

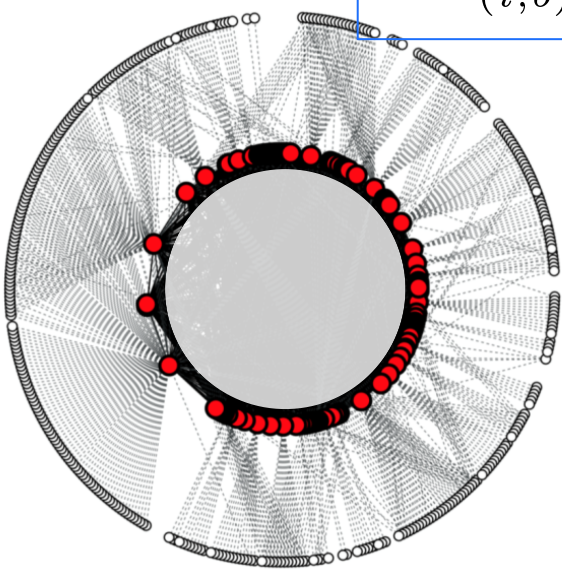
Details

$$N = I + \boxed{E} = \sum_{i \in C, j \in C} \left(A_{ij} - \frac{k_i k_j}{2m}\right) s(\mathbf{x_i}, \mathbf{x_j} | \mathbf{w})$$

$$\boxed{- \sum_{\substack{i \in C, b \in B \\ (i,b) \in \mathcal{E}}} \left(1 - \min\left(1, \frac{k_i k_b}{2m}\right)\right) s(\mathbf{x_i}, \mathbf{x_b} | \mathbf{w})}$$

1

external separability



The measure of Normality

Details

- Given an attributed subgraph, can we find the attribute weights?

$$N(C) = \sum_{\substack{i \in C, j \in C, \\ i \neq j}} \left(A_{ij} - \frac{k_i k_j}{2m}\right) sim_{\mathbf{w}}(\mathbf{x_i}, \mathbf{x_j}) - \sum_{\substack{i \in C, b \in B \\ (i,b) \in \mathcal{E}}} \left(1 - \min(1, \frac{k_i k_b}{2m})\right) sim_{\mathbf{w}}(\mathbf{x_i}, \mathbf{x_b})$$

1

$$\arg\max_{\mathbf{w}} \ \mathbf{w}^T \cdot \left[\sum_{\substack{i \in C, j \in C, \\ i \neq j}} \left(A_{ij} - \frac{k_i k_j}{2m}\right)(\mathbf{x_i} \odot \mathbf{x_j}) - \sum_{\substack{i \in C, b \in B \\ (i,b) \in \mathcal{E}}} \left(1 - \min(1, \frac{k_i k_b}{2m})\right)(\mathbf{x_i} \odot \mathbf{x_b}) \right]$$

latent

2

Optimizing Normality

Details

$$N = I + E = \sum_{i\in C, j\in C} \left(A_{ij} - \frac{k_i k_j}{2m}\right) s(\mathbf{x_i}, \mathbf{x_j} | \mathbf{w}) - \sum_{\substack{i\in C, b\in B \\ (i,b)\in\mathcal{E}}} \left(1 - \min(1, \frac{k_i k_b}{2m})\right) s(\mathbf{x_i}, \mathbf{x_b} | \mathbf{w})$$

1

$$\max_{\mathbf{w_C}} \quad \mathbf{w_C}^T \cdot \left[\sum_{i\in C, j\in C} \left(A_{ij} - \frac{k_i k_j}{2m}\right) s(\mathbf{x_i}, \mathbf{x_j}) - \sum_{\substack{i\in C, b\in B \\ (i,b)\in\mathcal{E}}} \left(1 - \min(1, \frac{k_i k_b}{2m})\right) s(\mathbf{x_i}, \mathbf{x_b}) \right]$$

2

$$\max_{\mathbf{w_C}} \quad \mathbf{w_C}^T \cdot (\hat{\mathbf{x}}_I + \hat{\mathbf{x}}_E)$$

$$\text{s.t.} \quad \|\mathbf{w_C}\|_p = 1, \;\; \mathbf{w_C}(f) \geq 0, \; \forall f = 1 \ldots d$$

3

Optimizing Normality

Details

$$\max_{\mathbf{w}_C} \quad {\mathbf{w}_C}^T \cdot (\underbrace{\hat{\mathbf{x}}_I + \hat{\mathbf{x}}_E}_{\mathbf{x}})$$

$$\text{s.t.} \quad \|\mathbf{w}_C\|_p = 1, \quad \mathbf{w}_C(f) \geq 0, \ \forall f = 1 \ldots d$$

$p = 1$: $\mathbf{w}_C(f) = 1$ **one** attribute f with largest **x**

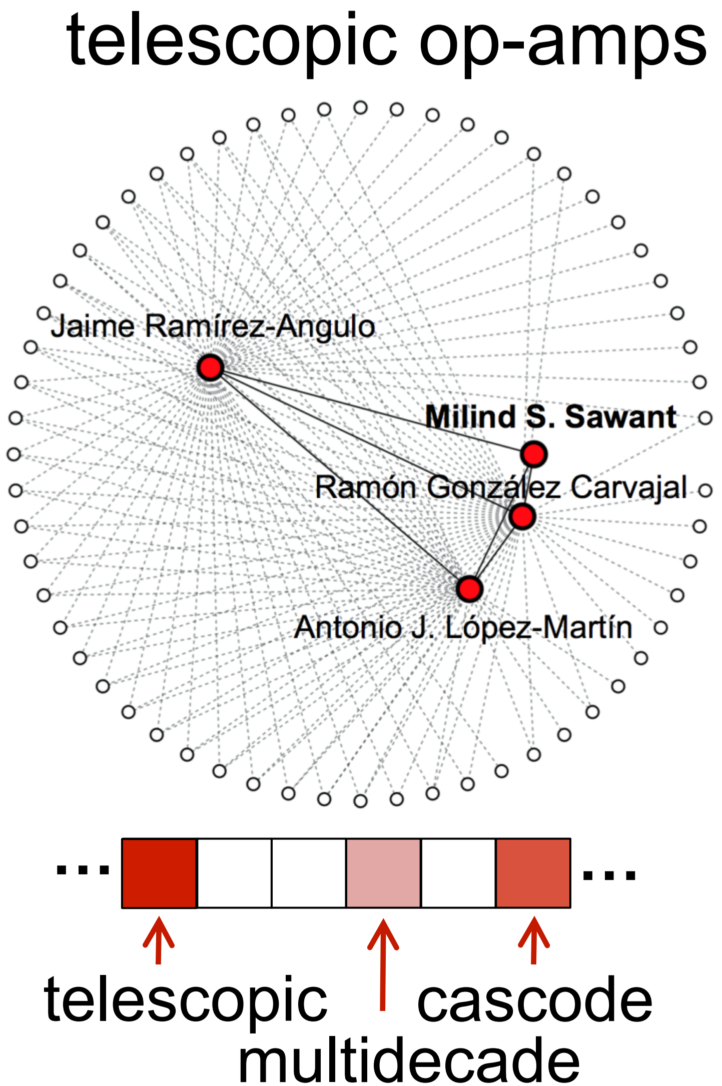
$p = 2$: $\mathbf{w}_C(f) = \frac{\mathbf{x}(f)}{\sqrt{\sum_{\mathbf{x}(i)>0} \mathbf{x}(i)^2}}$ **all** f with positive **x**

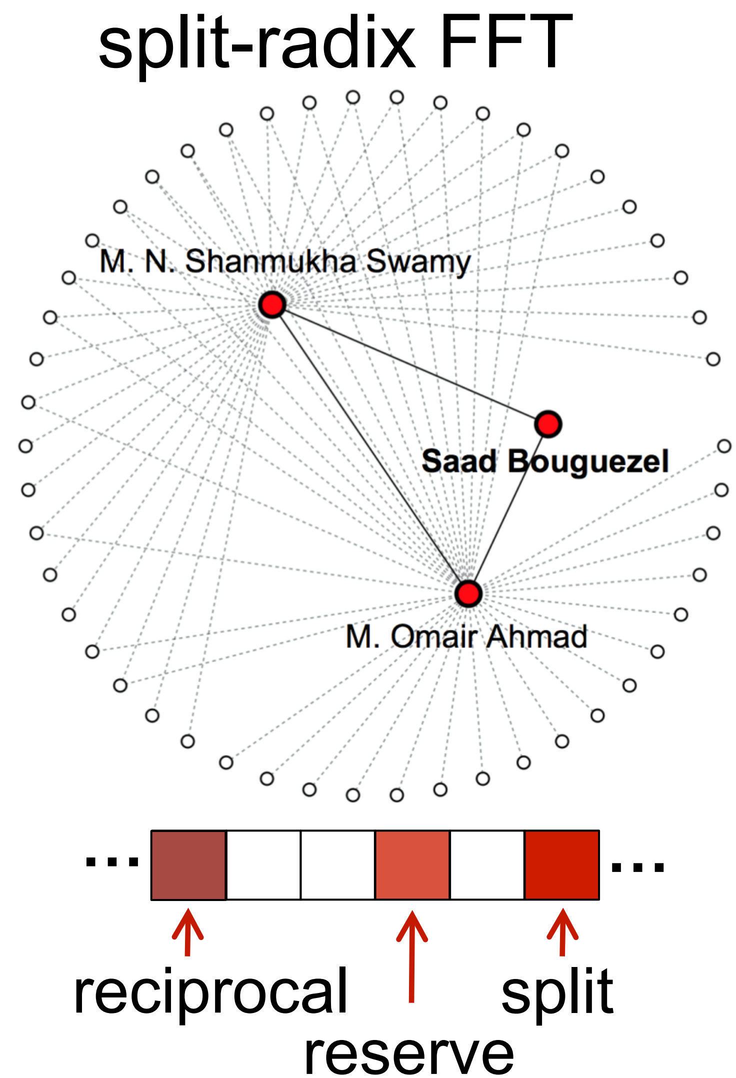
Normality becomes $N = {\mathbf{w}_C}^T \cdot \mathbf{x} = \|\mathbf{x}_+\|_2$

Linear in number of attributes!

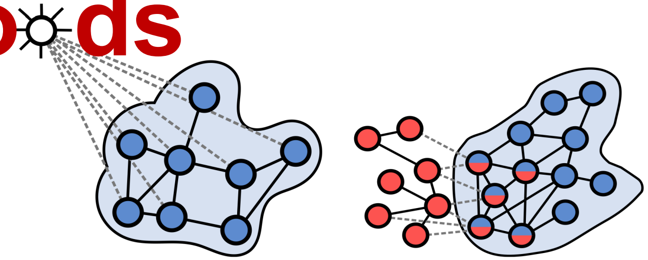
when $p = 1$, $N \in [-1, 1]$ $\qquad$ $N \in [-1, \|\mathbf{x}_+\|_2]$ when $p = 2$.

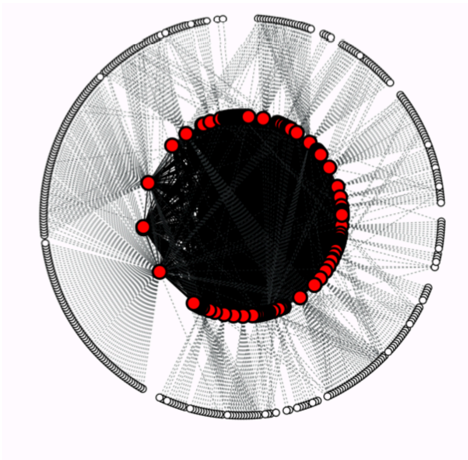
Illustrative examples





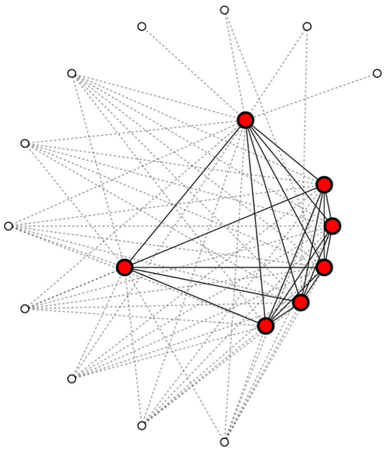
Example neighborhoods





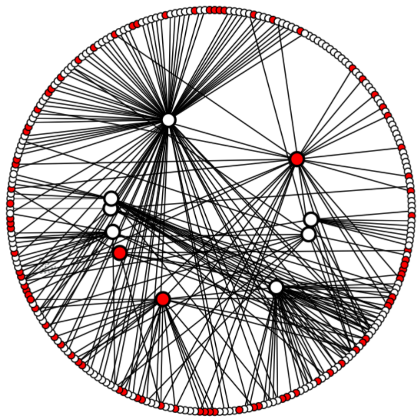
DBLP

$L_1 = 0.979,\ L_2 = 2.17$



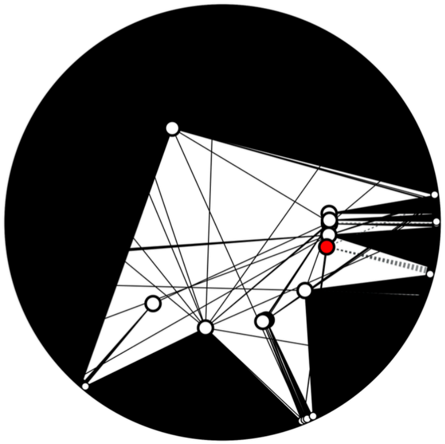
Twitter

$L_1 = 0.724,\ L_2 = 1.10$



Google+

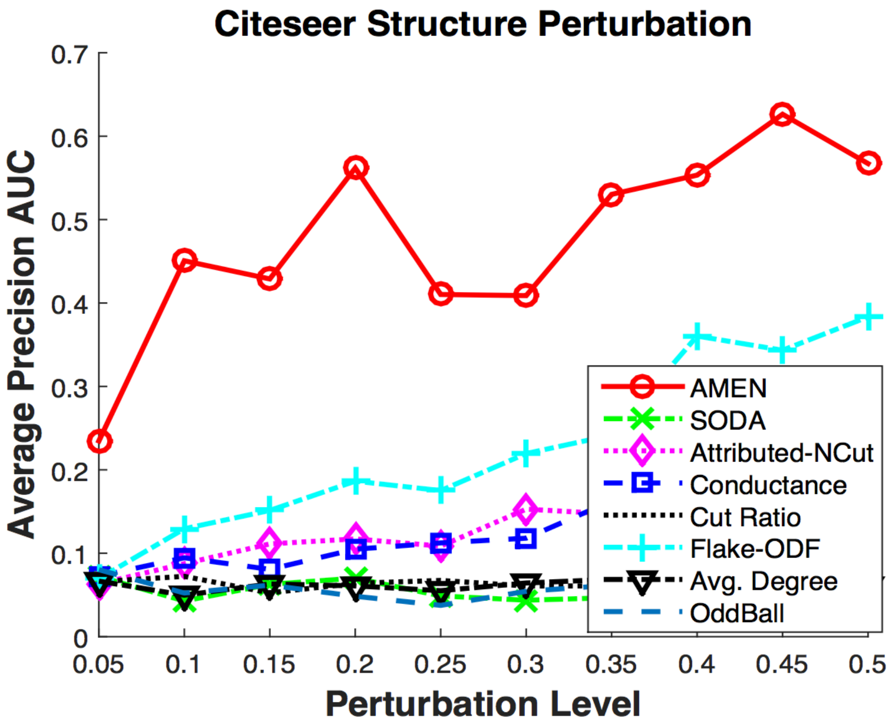
$L_1 = L_2 = -0.873$

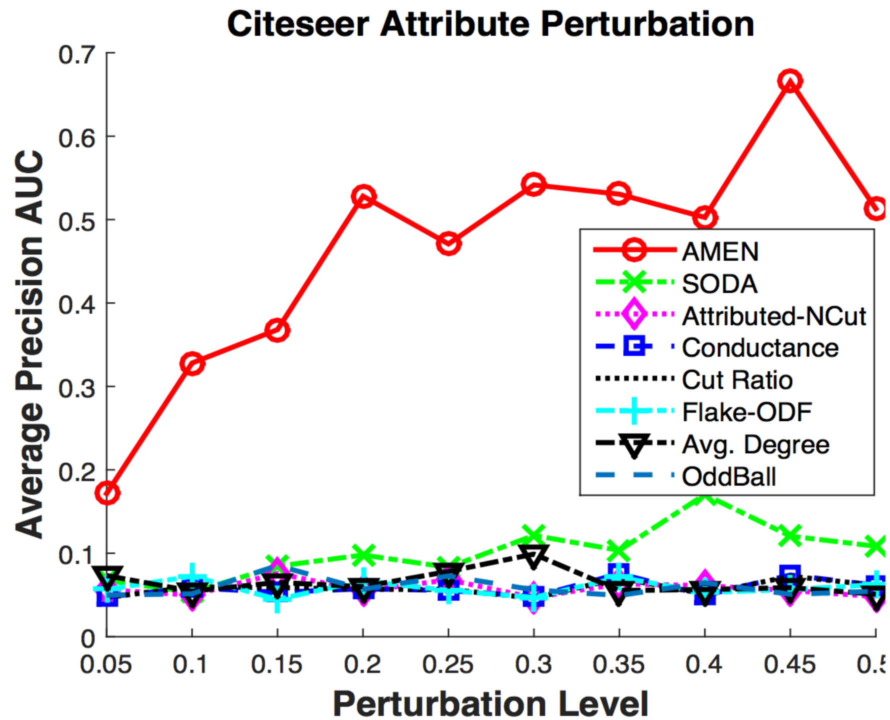


Citeseer

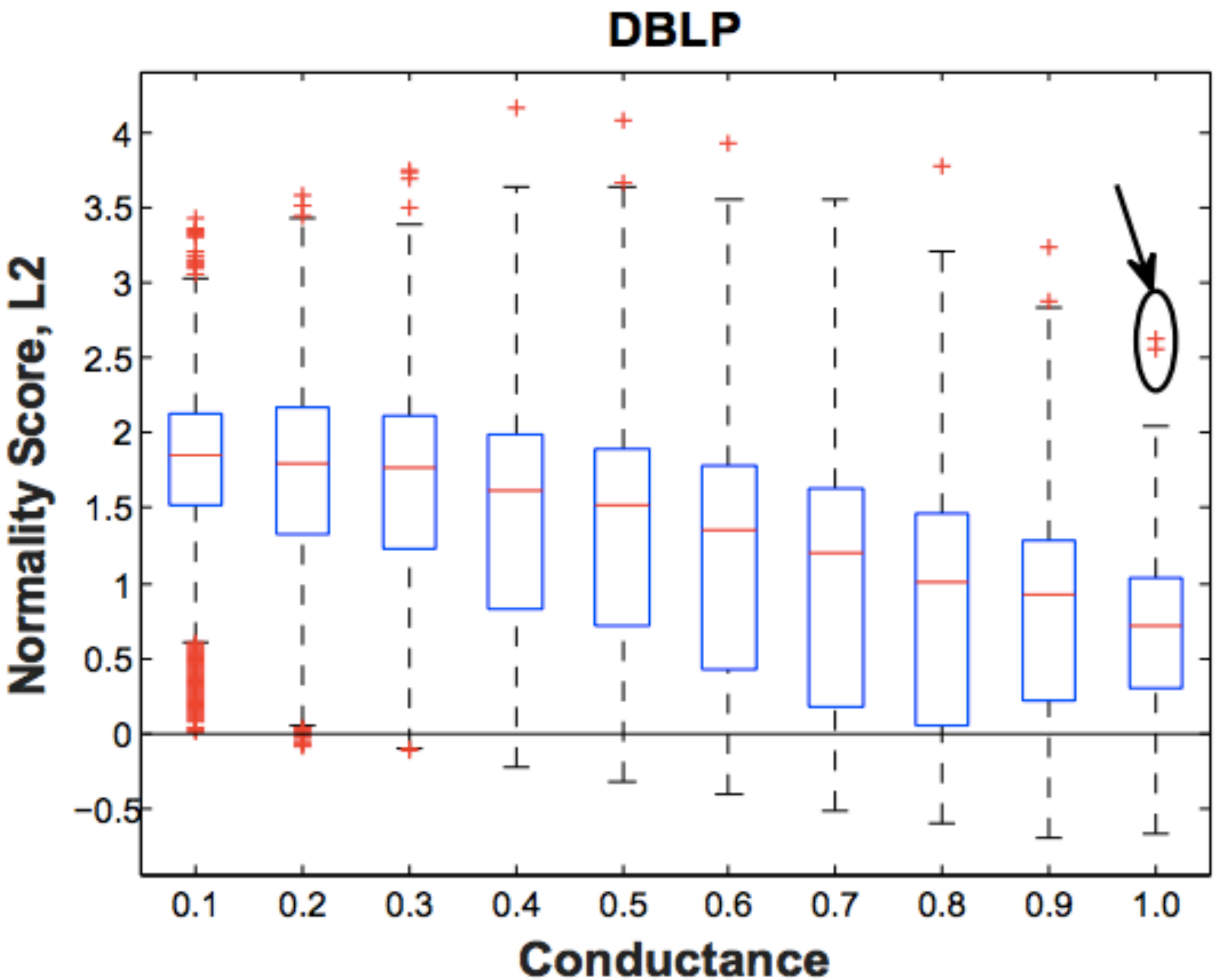
$L_1 = L_2 = -0.956$

Anomaly detection: Perturbed data

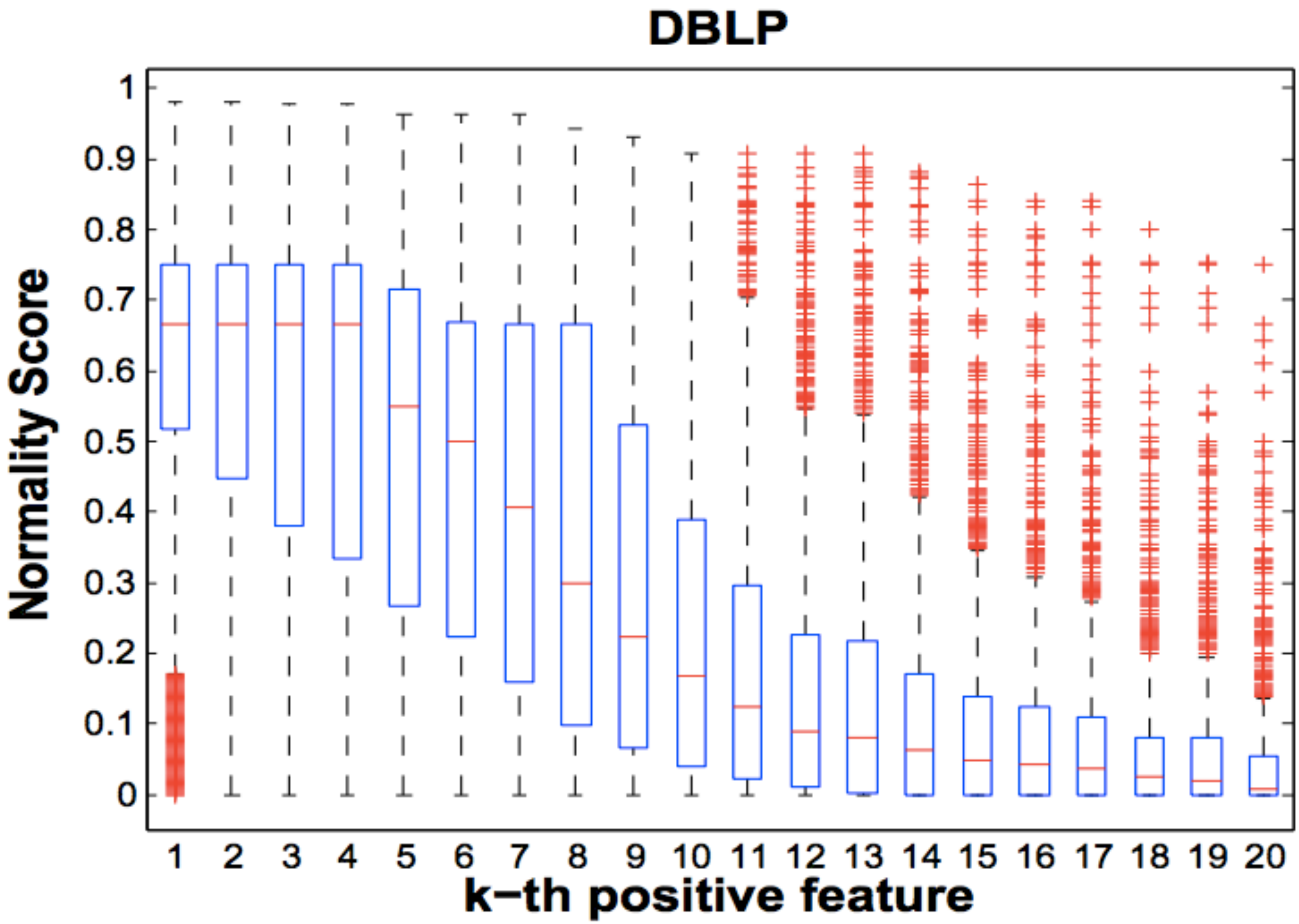




Normality vs Conductance, DBLP



Attribute distribution, DBLP

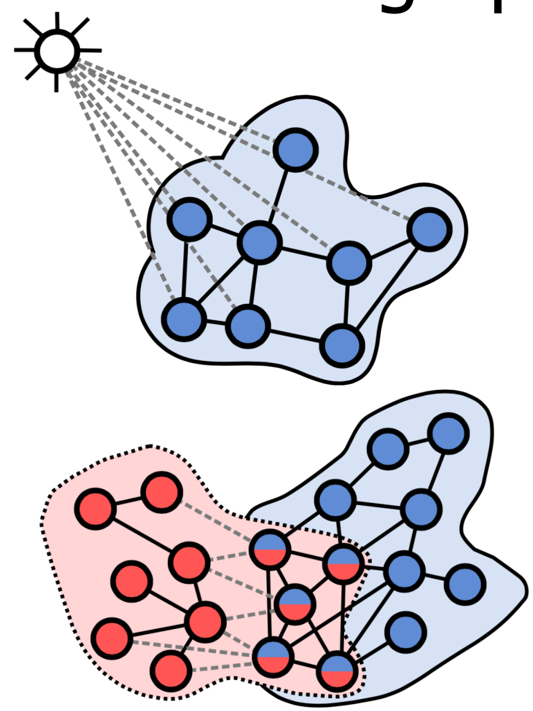


Summary

A new quality measure for attributed subgraphs

normality considers:

- internal + boundary
- structure + attributes
- subgraph **focus**
- "*exoneration*"



Automatic inference of **focus**

- via **normality** maximization
- unsupervised
- linear in #attributes

Paper, code, data

- http://www.perozzi.net/projects/amen/

Bryan Perozzi

Overview

» About Me
» Research Interests
» Selected Publications
» Honors and Awards
» Press Coverage

Publications

» Conference & Journal
» Workshop & Poster

Projects

» Anomaly Detection in Attributed Graphs

Anomaly Ranking of Attributed Neighborhoods

Bryan Perozzi, Leman Akoglu
May 9, 2016

Awards: **Best Paper Runner-up, SDM'16!**

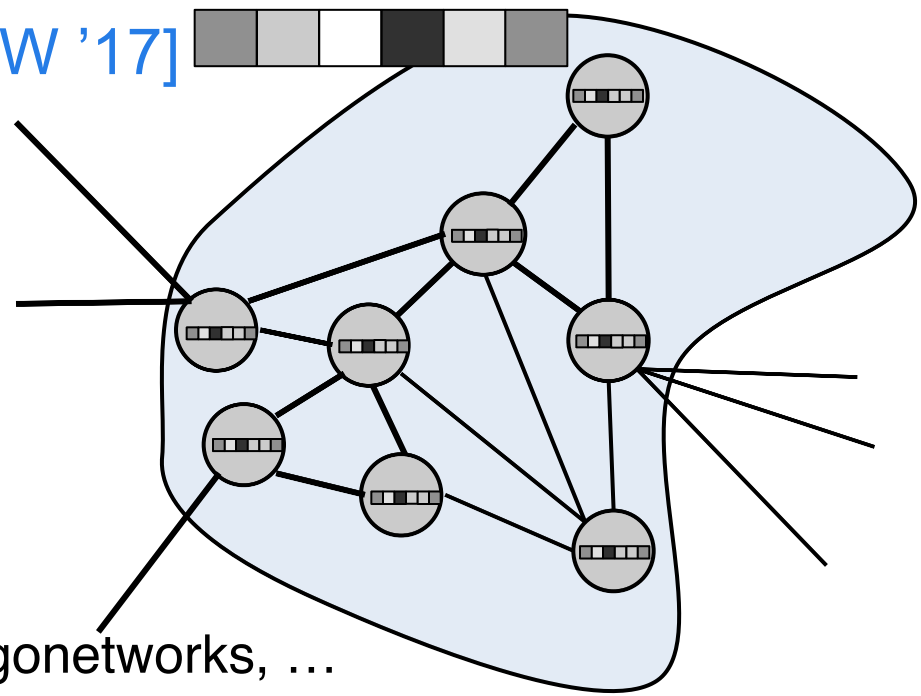
Overview

Given a graph with node attributes, what neighborhoods are anomalous? To answer this question, one needs a quality score that utilizes both structure and attributes. Popular existing measures either quantify the structure only and ignore the attributes (e.g., ... and ignore the cross-edges at the boundary (e.g., density).

Scalable Anomaly Ranking of Attributed Neighborhoods
Bryan Perozzi and Leman Akoglu *SIAM SDM 2016.*

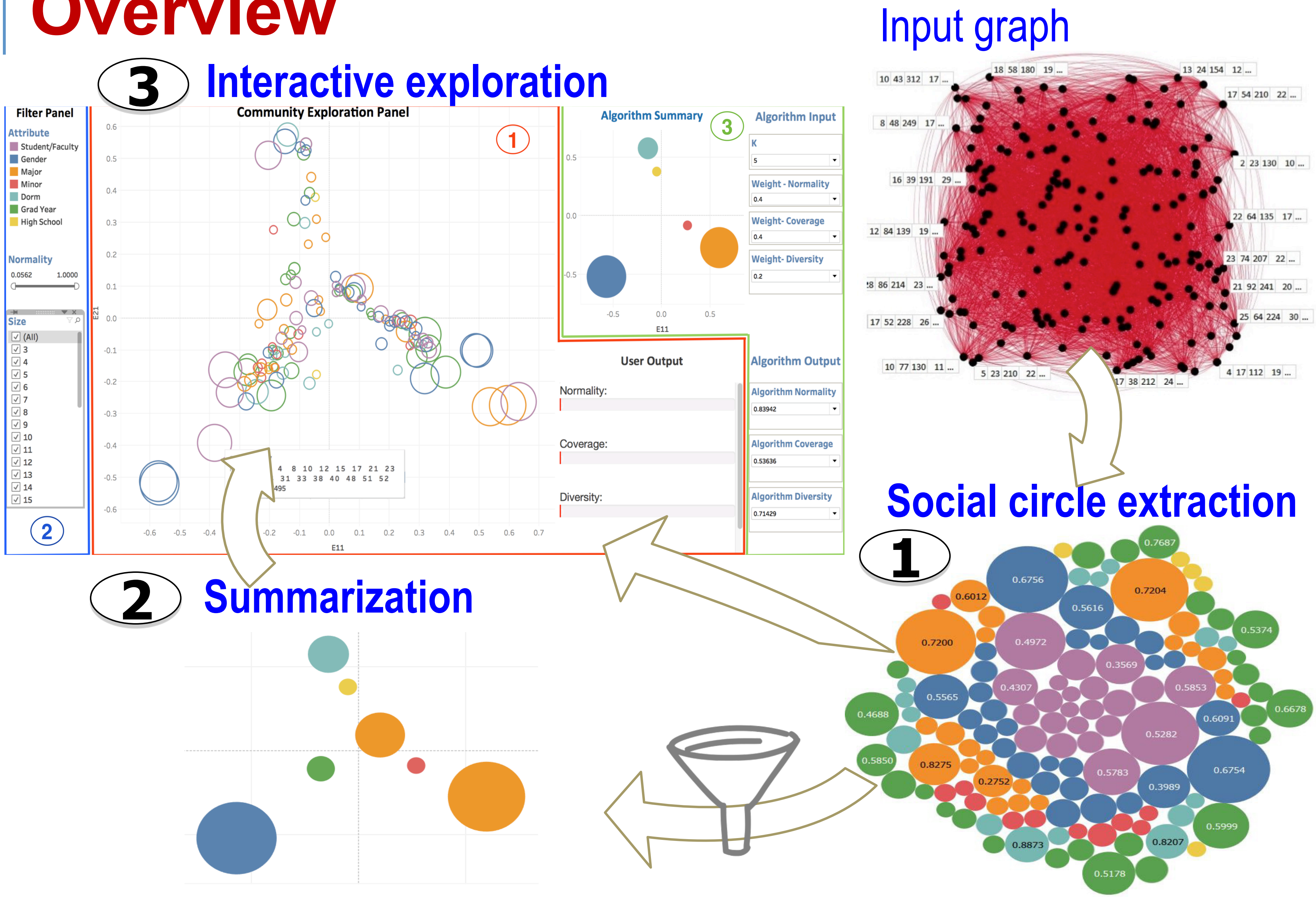
This talk

- Attributed (sub)graphs*
  - Subgraphs [SIAM SDM'16]
  - Summarization [ACM TKDD'18]
  - Comparisons [WWW '17]



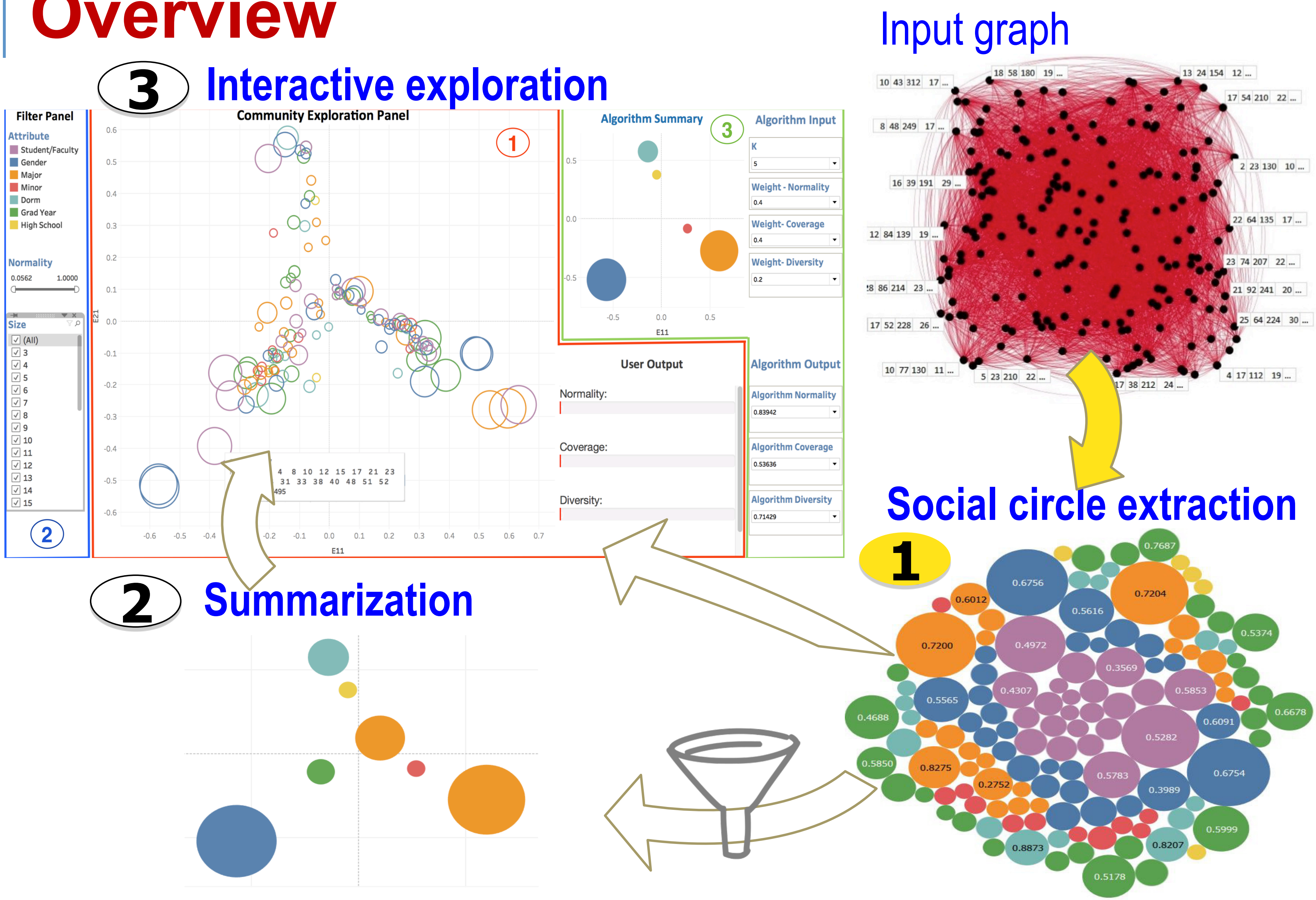
* social circles, communities, egonetworks, …

Overview



Input graph
3 Interactive exploration
Filter Panel
Attribute
Student/Faculty
Gender
Major
Minor
Dorm
Grad Year
High School
Normality
0.0562
1.0000
Size
(All)
Community Exploration Panel
Algorithm Summary
Algorithm Input
K
5
Weight - Normality
0.4
Weight- Coverage
0.4
Weight- Diversity
0.2
User Output
Normality:
Coverage:
Diversity:
Algorithm Output
Algorithm Normality
0.83942
Algorithm Coverage
0.53636
Algorithm Diversity
0.71429
Social circle extraction
1
2 Summarization

Overview



Input graph
3 Interactive exploration
Filter Panel
Attribute
Student/Faculty
Gender
Major
Minor
Dorm
Grad Year
High School
Normality
0.0562
1.0000
Size
(All)
3
4
5
6
7
8
9
10
11
12
13
14
15
2
Community Exploration Panel
1
E21
E11
Algorithm Summary
3
Algorithm Input
K
5
Weight - Normality
0.4
Weight- Coverage
0.4
Weight- Diversity
0.2
User Output
Normality:
Coverage:
Diversity:
Algorithm Output
Algorithm Normality
0.83942
Algorithm Coverage
0.53636
Algorithm Diversity
0.71429
Social circle extraction
1
2 Summarization

Extracting Social Circles

- a GRASP (Greedy Randomized Adaptive Search Procedure) approach [Feo & Resende '95]

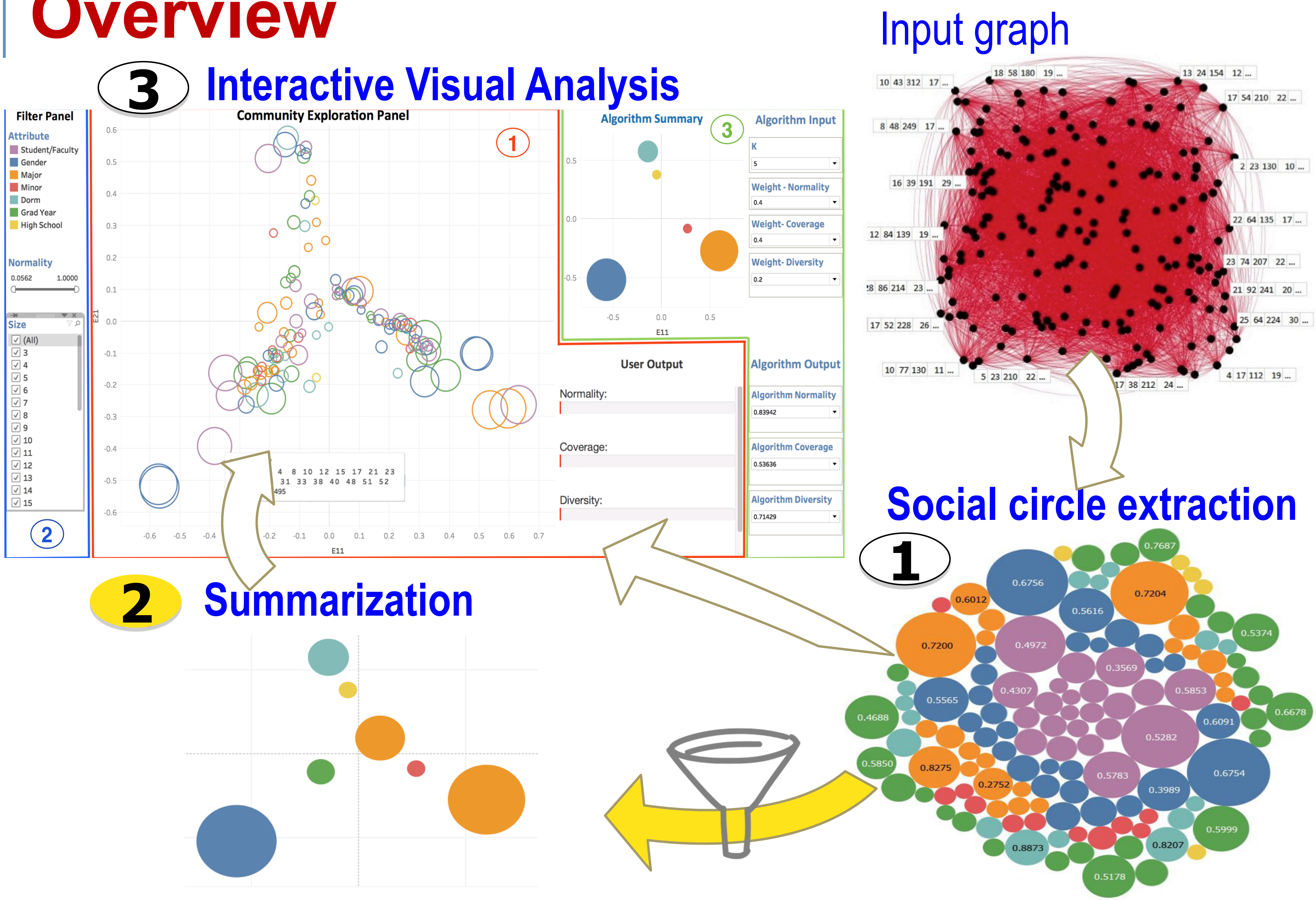
Algorithm 1 EXTRACTATTRIBUTEDSOCIALCIRCLES

Input: $G = (\mathcal{V}, \mathcal{E}, \mathcal{A})$, node attribute vectors $\mathbf{x}_{u \in \mathcal{V}}$, T_{max}, α
Output: set of extracted communities $\mathcal{C}$

```
1: C := ∅
2: for each u ∈ V do
3:     for t = 1 : T_max do
4:         S := CONSTRUCTION(u, G, α)
5:         C := C ∪ LOCALSEARCH(S, G)
6:     end for
7: end for
8: return C
```

- note: one focus attribute per circle

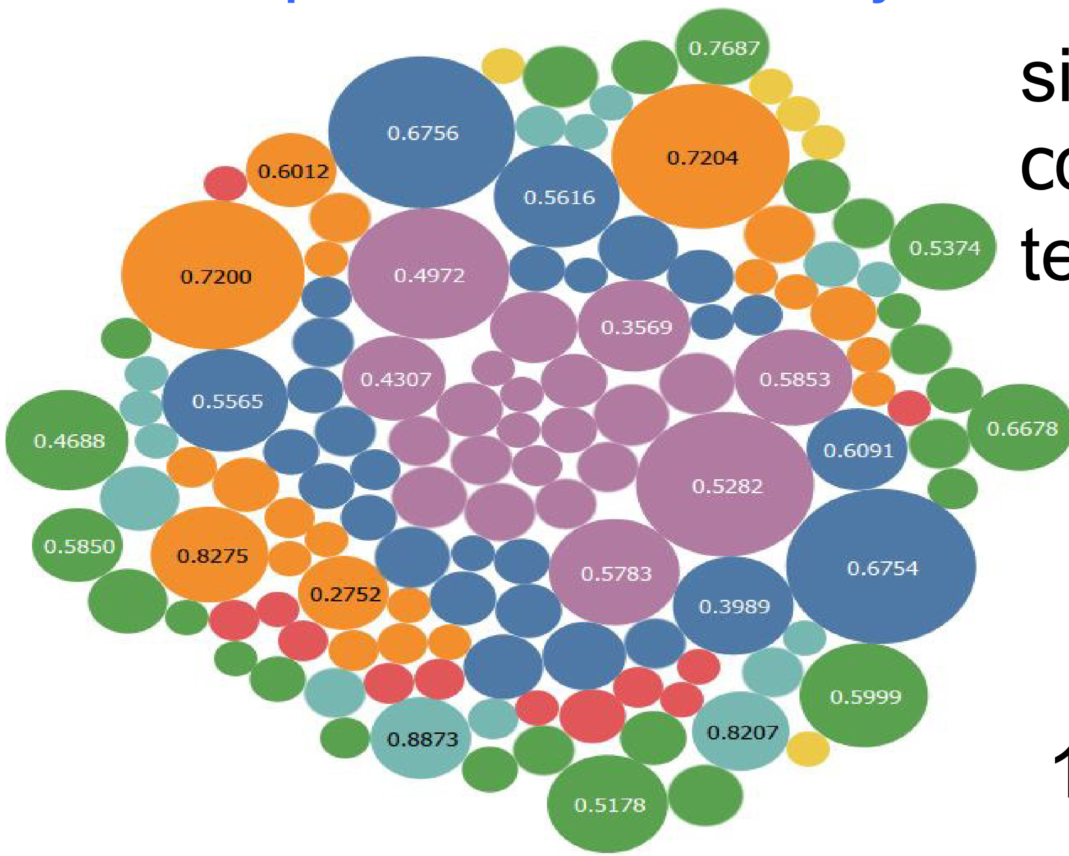
Overview



Input graph
3
Interactive Visual Analysis
Filter Panel
Attribute
Student/Faculty
Gender
Major
Minor
Dorm
Grad Year
High School
Normality
0.0562
1.0000
Size
(All)
2
Community Exploration Panel
1
E21
E11
Algorithm Summary
3
Algorithm Input
K
5
Weight - Normality
0.4
Weight- Coverage
0.4
Weight- Diversity
0.2
User Output
Normality:
Coverage:
Diversity:
Algorithm Output
Algorithm Normality
0.83942
Algorithm Coverage
0.53636
Algorithm Diversity
0.71429
Social circle extraction
1
2
Summarization

Summarization

- Social circles: what size, quality and focus?
  - Attempt: visual summary



size $\propto$ #nodes
color: 'focus'
text: normality

125 circles!

  - does not reflect overlap between circles!

Summarization

Details

- Want a summary (a few circles):
  - high normality
  - well-“cover” the graph
  - diverse in ‘focus’

$$\max_{\substack{S \subseteq \mathcal{C} \\ |S| = K}} \quad f(S) = \alpha \; avgnorm(S) + \beta \; cov(S) + (1 - \alpha - \beta) \; div(S)$$

$$= \alpha \frac{\sum_{C \in S} N(C)}{K} + \beta \frac{|\bigcup_{C \in S} C|}{n} + (1 - \alpha - \beta) \frac{|\bigcup_{C \in S} \mathcal{A}(C)|}{d}$$

$0 \leq \alpha, \beta \leq 1$ can be interactively adjusted by users

Summarization

Details

$$\max_{\substack{S \subseteq \mathcal{C} \\ |S|=K}} \quad f(S) = \alpha \underbrace{\frac{\sum_{C \in S} N(C)}{K}}_{\text{avg. normality}} + \beta \underbrace{\frac{|\bigcup_{C \in S} C|}{n}}_{\text{coverage}} + (1-\alpha-\beta) \underbrace{\frac{|\bigcup_{C \in S} \mathcal{A}(C)|}{d}}_{\text{diversity}}$$

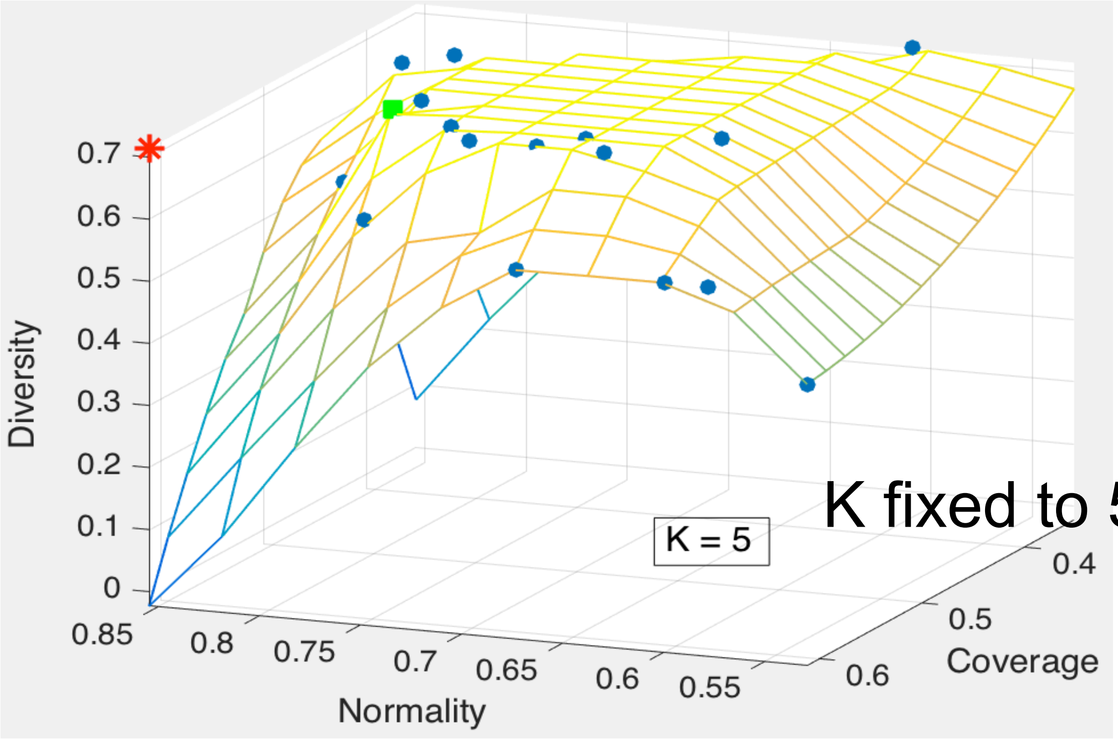
- Provided K, n, d (denominators) fixed, easy to show that $f : 2^{\mathcal{C}} \rightarrow \mathbb{R}_+$ is
  - non-negative
  - monotonic: $A \subseteq B \subseteq \mathcal{C},\ f(A) \leq f(B)$
  - submodular: for every $A \subseteq B \subseteq \mathcal{C}$ and $C \in \mathcal{C} \backslash B$,

$$f(A \cup \{C\}) - f(A) \geq f(B \cup \{C\}) - f(B)$$

→ The "next-best" greedy algorithm: at least 63% of the objective value f(·) of the *optimum* set.

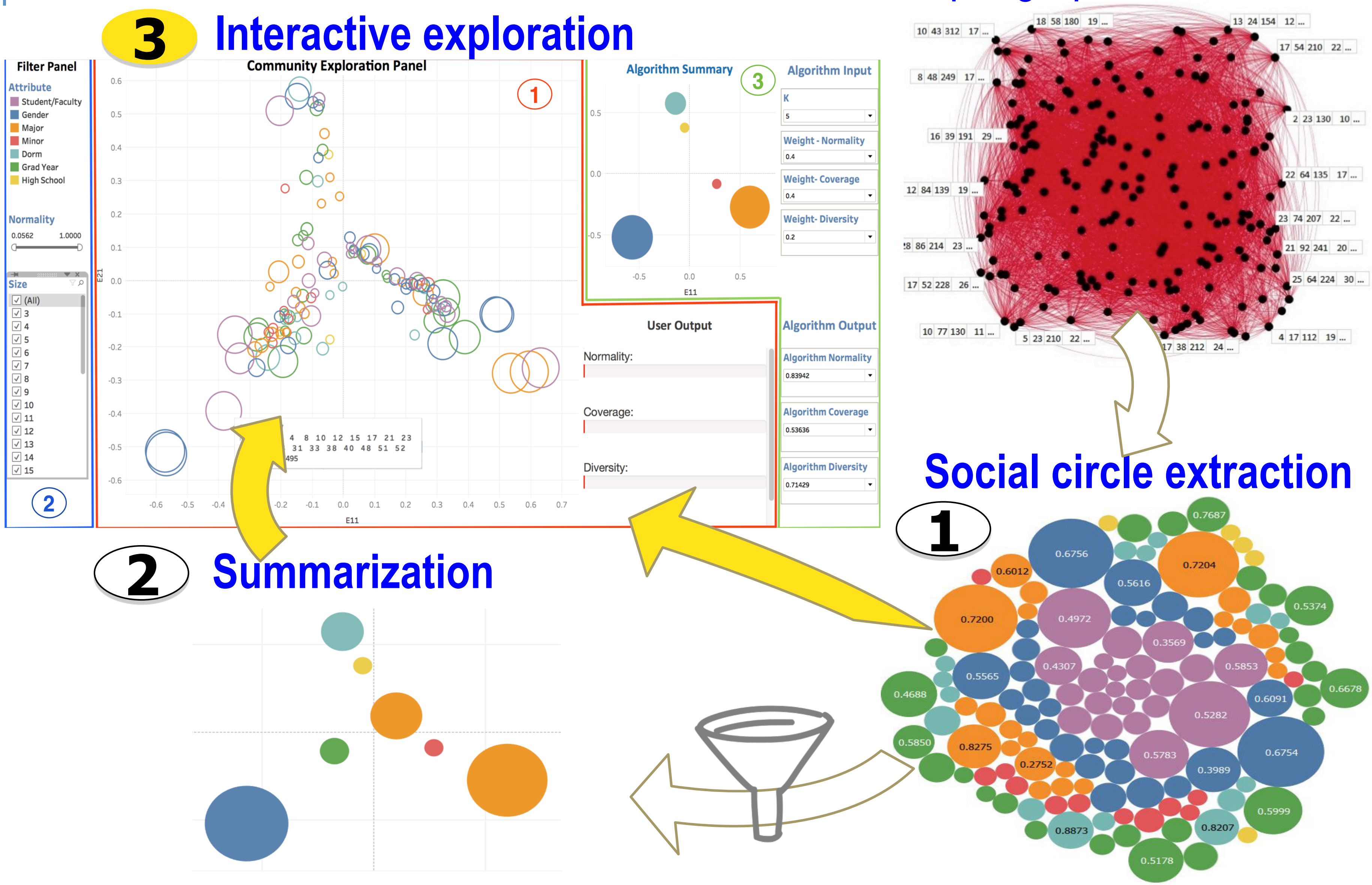
Summarization

- surface formed by various parameter combinations $(\alpha, \beta, 1-\alpha-\beta)$ (blue dots)
  - (green) square around the "knee": a good trade-off between quality, coverage, and diversity



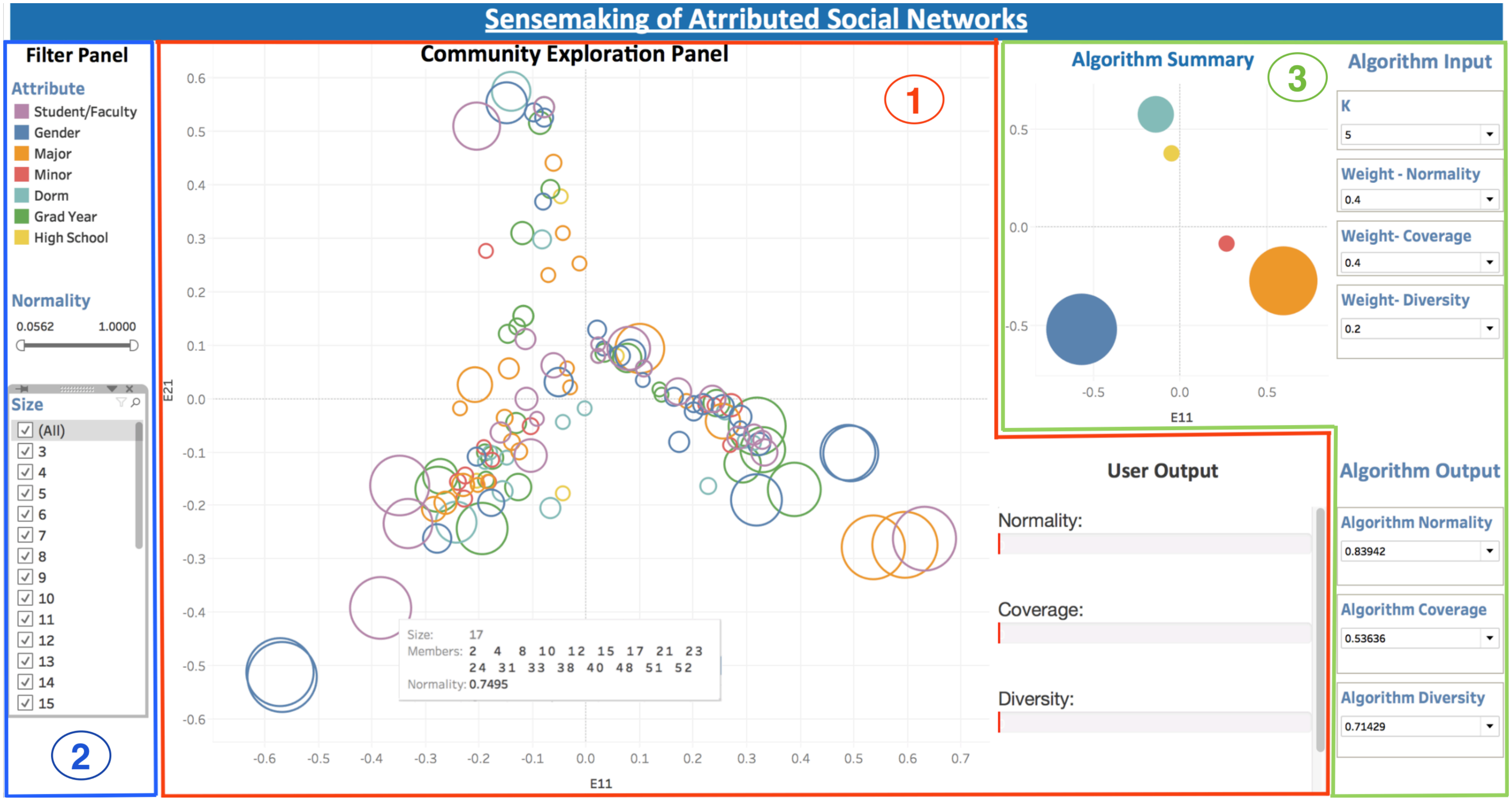
K fixed to 5

Overview

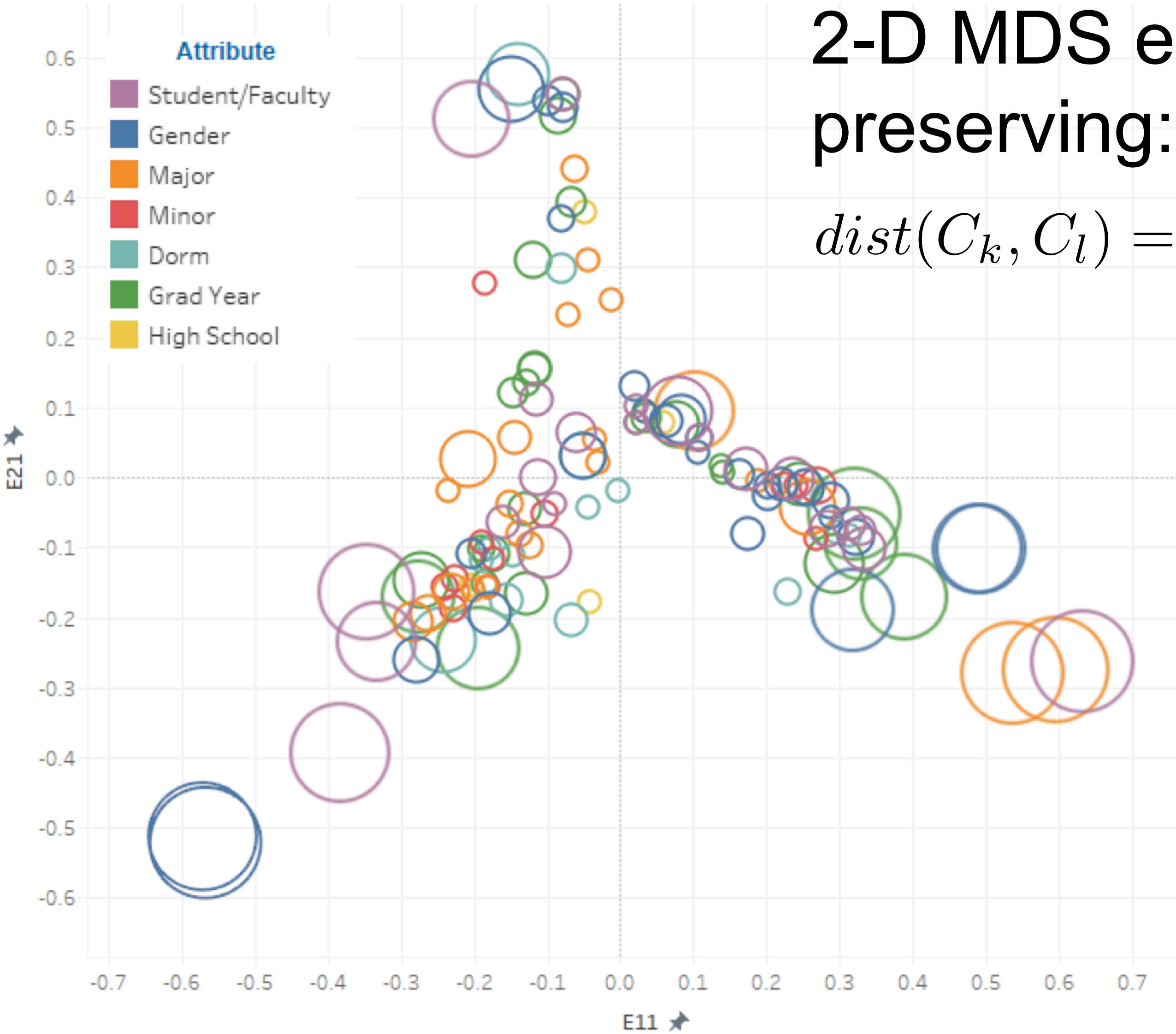


3 Interactive exploration
Input graph
Filter Panel
Attribute
Student/Faculty
Gender
Major
Minor
Dorm
Grad Year
High School
Normality
0.0562
1.0000
Size
(All)
Community Exploration Panel
E11
E21
Algorithm Summary
Algorithm Input
K
Weight - Normality
Weight- Coverage
Weight- Diversity
User Output
Normality:
Coverage:
Diversity:
Algorithm Output
Algorithm Normality
Algorithm Coverage
Algorithm Diversity
Social circle extraction
1
2 Summarization

Interactive Visual Exploration & Summarization



Circle embedding



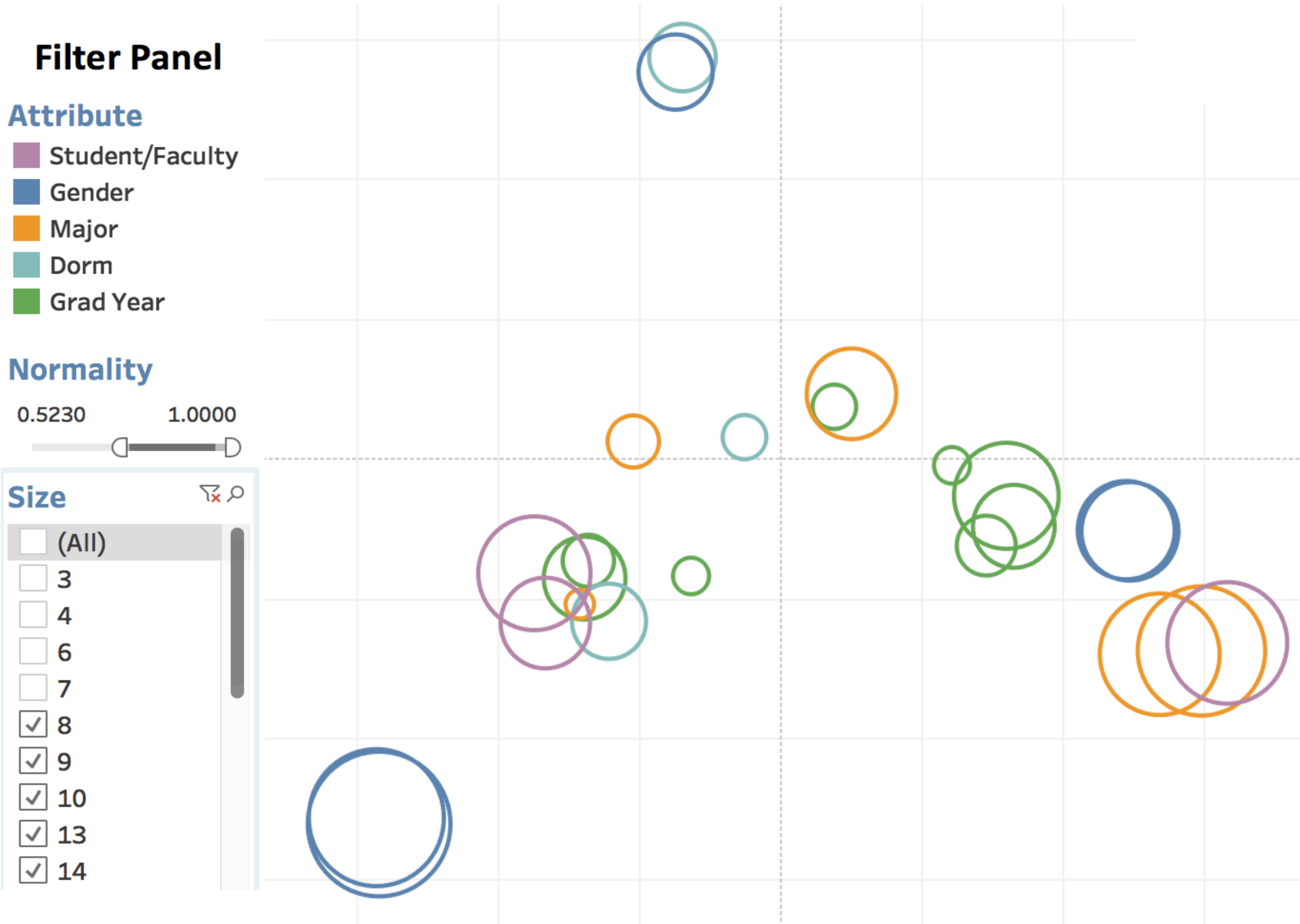
2-D MDS embedding preserving:

$$dist(C_k, C_l) = 1 - \frac{|C_k \cap C_l|}{\min(|C_k|, |C_l|)}$$

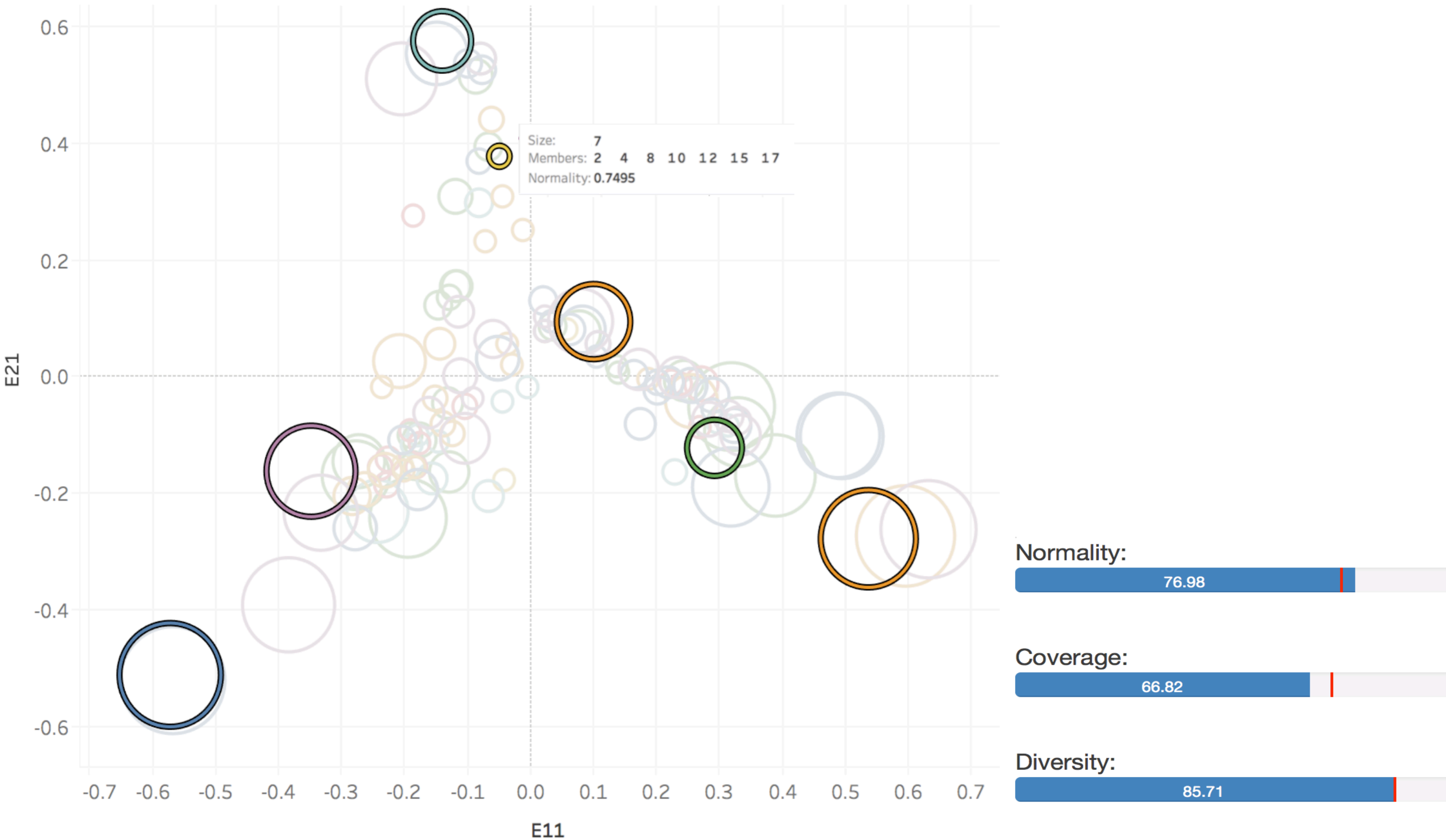
size $\propto$ #nodes

color: focus

Interaction: Filtering

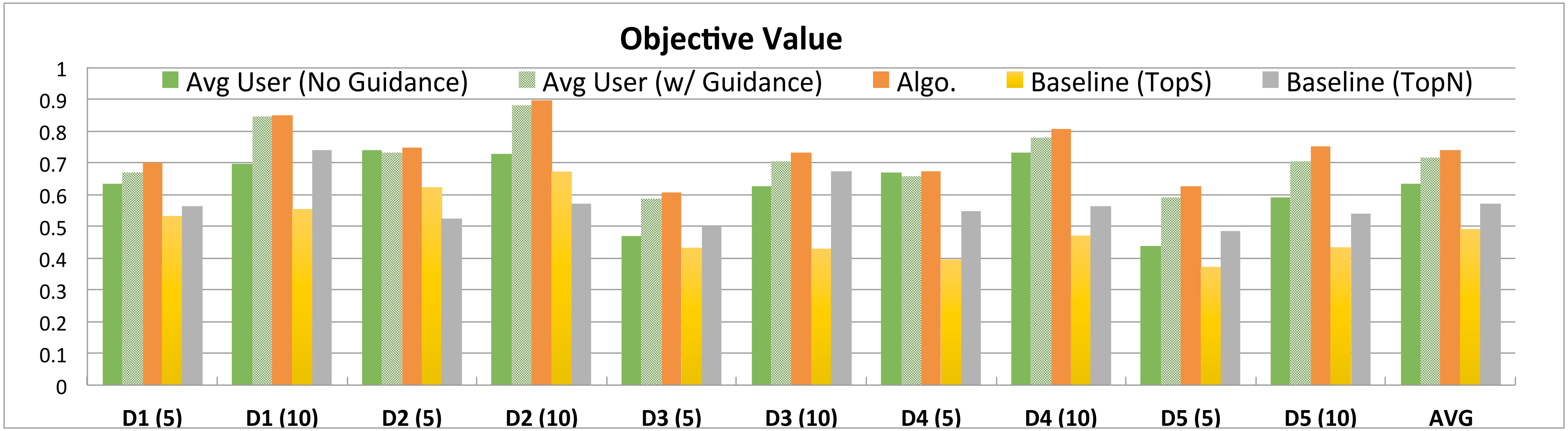


Interaction: Circle summarization



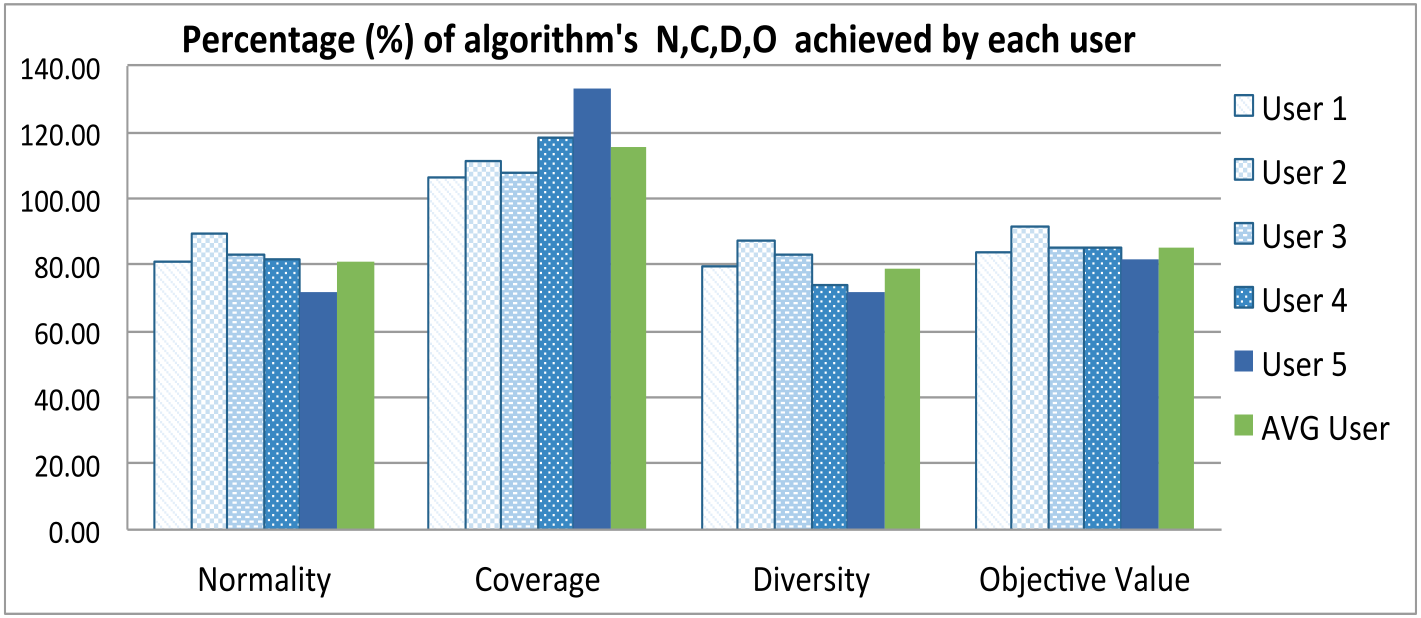
Evaluation

Q1) Summarization by visual exploration. *Does interactive visualization help users construct effective summaries, as compared to strawman baselines?*



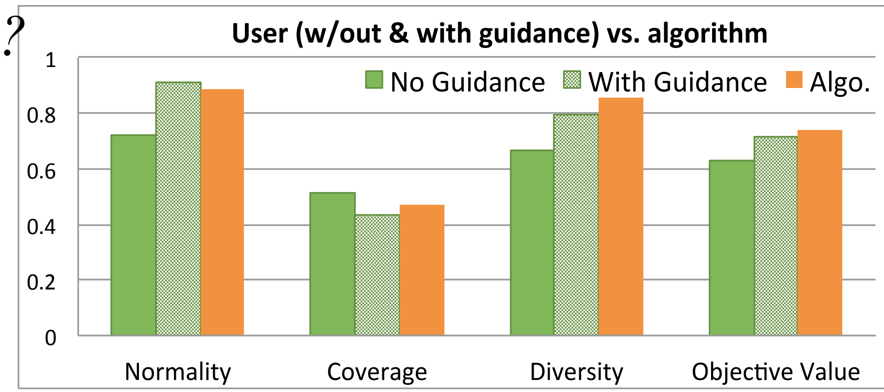
Evaluation

Q2) *How close do the summaries by users* ***without guidance*** *get to the algorithm results (in terms of normality, coverage, diversity, and overall objective value)?*



Evaluation

Q3) Alternative summarization by algorithmic guidance.
How much **guidance** *does our summarization algorithm provide users to derive alternative summaries and improve over their earlier results?*



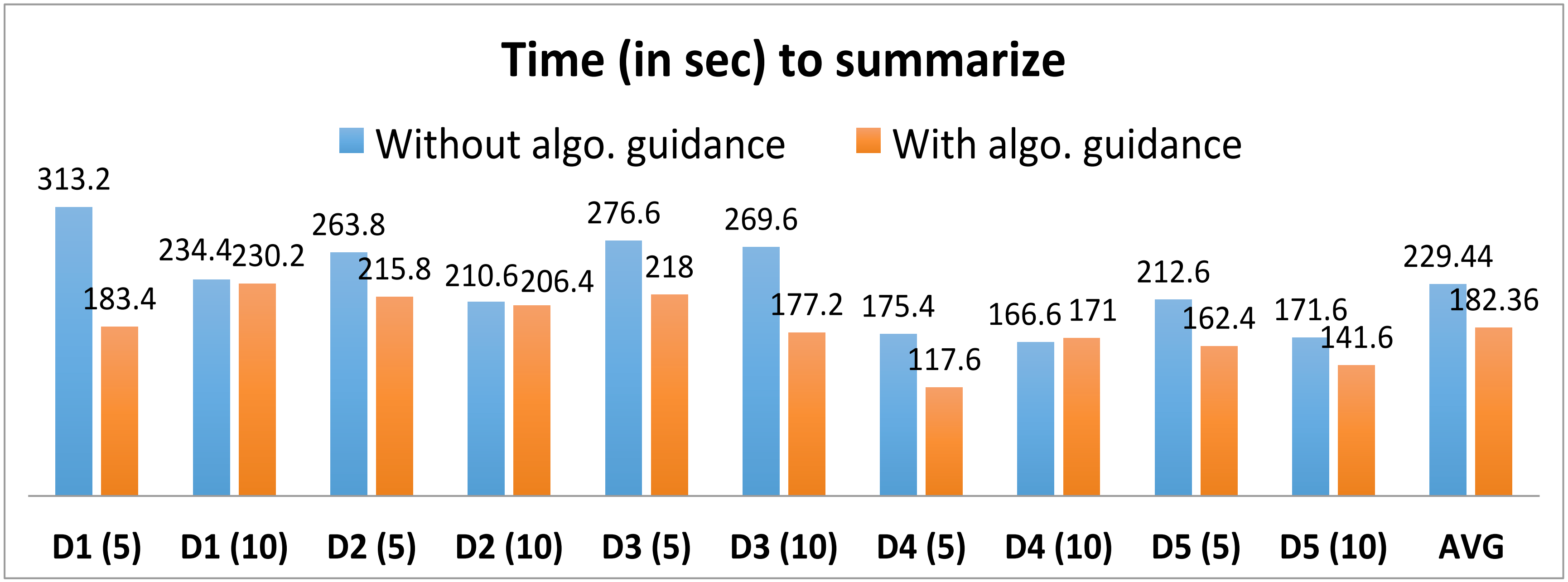
$$100\, O_{user}^{(after)} / O_{user}^{(before)}$$

PERCENT % IMPROVEMENT IN OBJECTIVE VALUE BY EACH USER ON EACH DATA/TASK **AFTER ALGORITHMIC GUIDANCE**.

	D1 (5)	D1 (10)	D2 (5)	D2 (10)	D3 (5)	D3 (10)	D4 (5)	D4 (10)	D5 (5)	D5 (10)	
User 1	112.59	156.44	99.53	114.31	129.89	130.58	92.20	106.17	170.86	121.08	**123.37**
User 2	91.79	118.14	87.56	102.86	99.19	112.31	92.66	100.00	107.39	117.97	**102.99**
User 3	101.60	112.95	101.30	120.73	140.15	101.75	85.78	96.60	199.57	142.96	**120.34**
User 4	103.98	104.18	100.85	140.65	103.76	105.94	116.86	124.73	110.13	109.13	**112.02**
User 5	117.61	124.02	102.70	129.06	169.17	117.77	105.06	106.17	113.34	109.65	**119.45**
Avg User	105.51	123.15	98.39	121.52	128.43	113.67	98.51	106.73	140.26	120.16	**115.63**

Evaluation

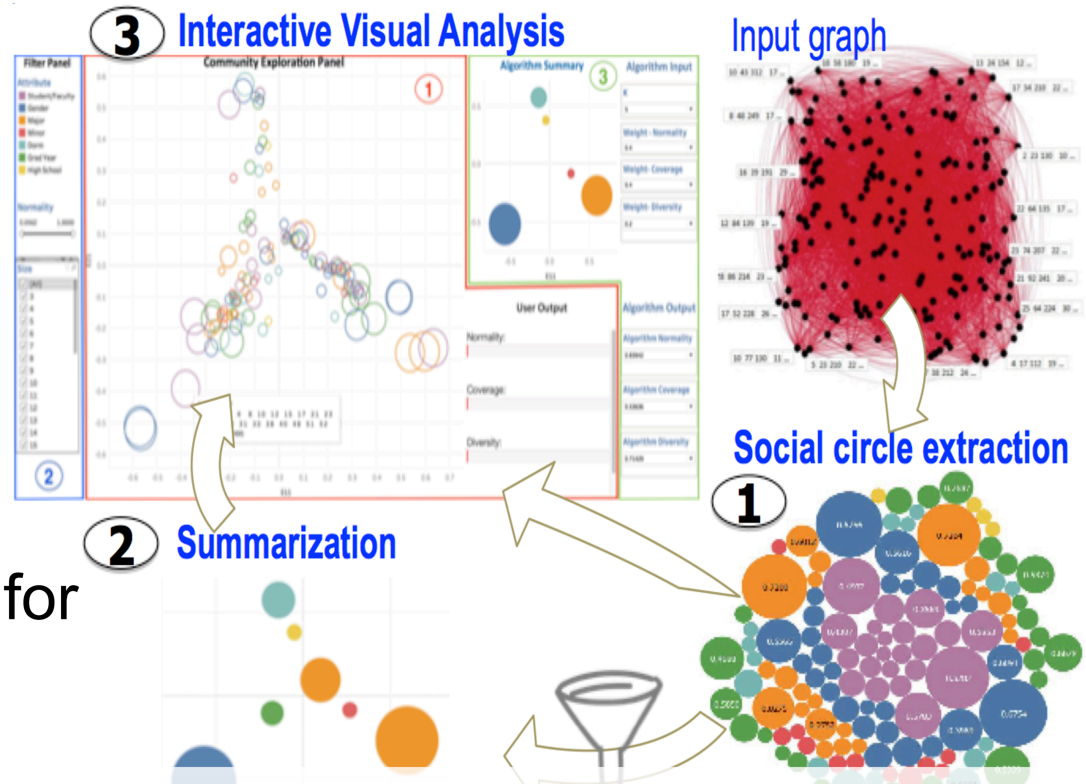
Q4) **Efficiency.** *How long does it take per user on average to construct (i) a summary without guidance, and (ii) alternative summary with guidance?*



Summary

- An end-to-end system for sensemaking of node-attributed networks

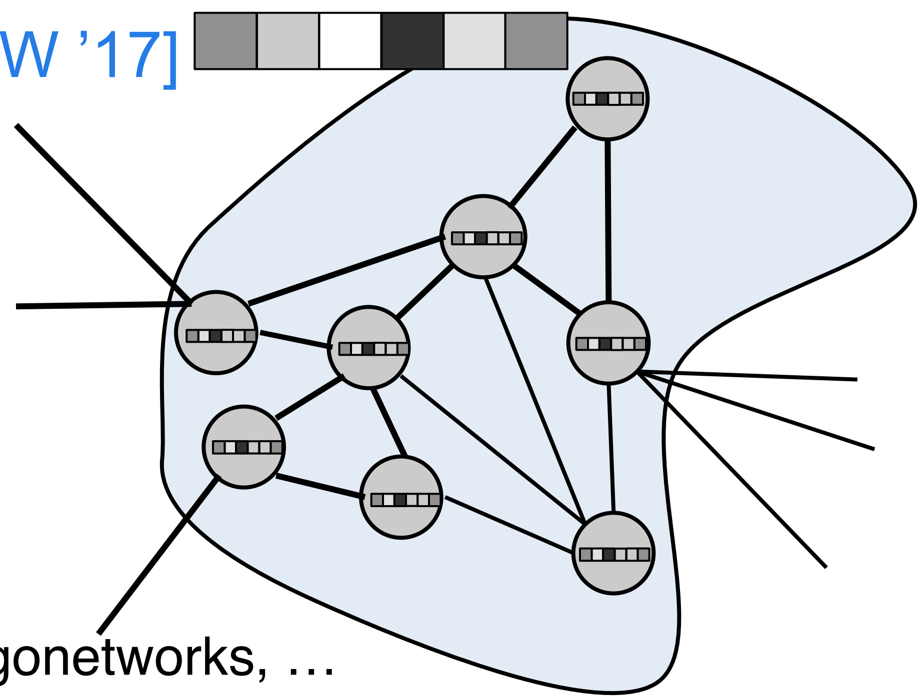
1. Circle extraction based on normality
2. Summarization wrt
   - quality,
   - coverage, and
   - diversity
3. Interactive interface for
   - exploration.



Discovering Communities and Anomalies in Attributed Graphs: Interactive Visual Exploration and Summarization *Bryan Perozzi and Leman Akoglu* ACM TKDD, 2018

This talk

- Attributed (sub)graphs*
  - Subgraphs [SIAM SDM'16]
  - Summarization [ACM TKDD'18]
  - Comparisons [WWW '17]



* social circles, communities, egonetworks, …

Comparing attributed (sub)graphs

- Motivating question:

Given a collection of attributed subgraphs
from different classes,
how can we discover the attributes that
characterize their differences?

- Hypothesis: subgraphs from different classes exhibit *different* focus attributes

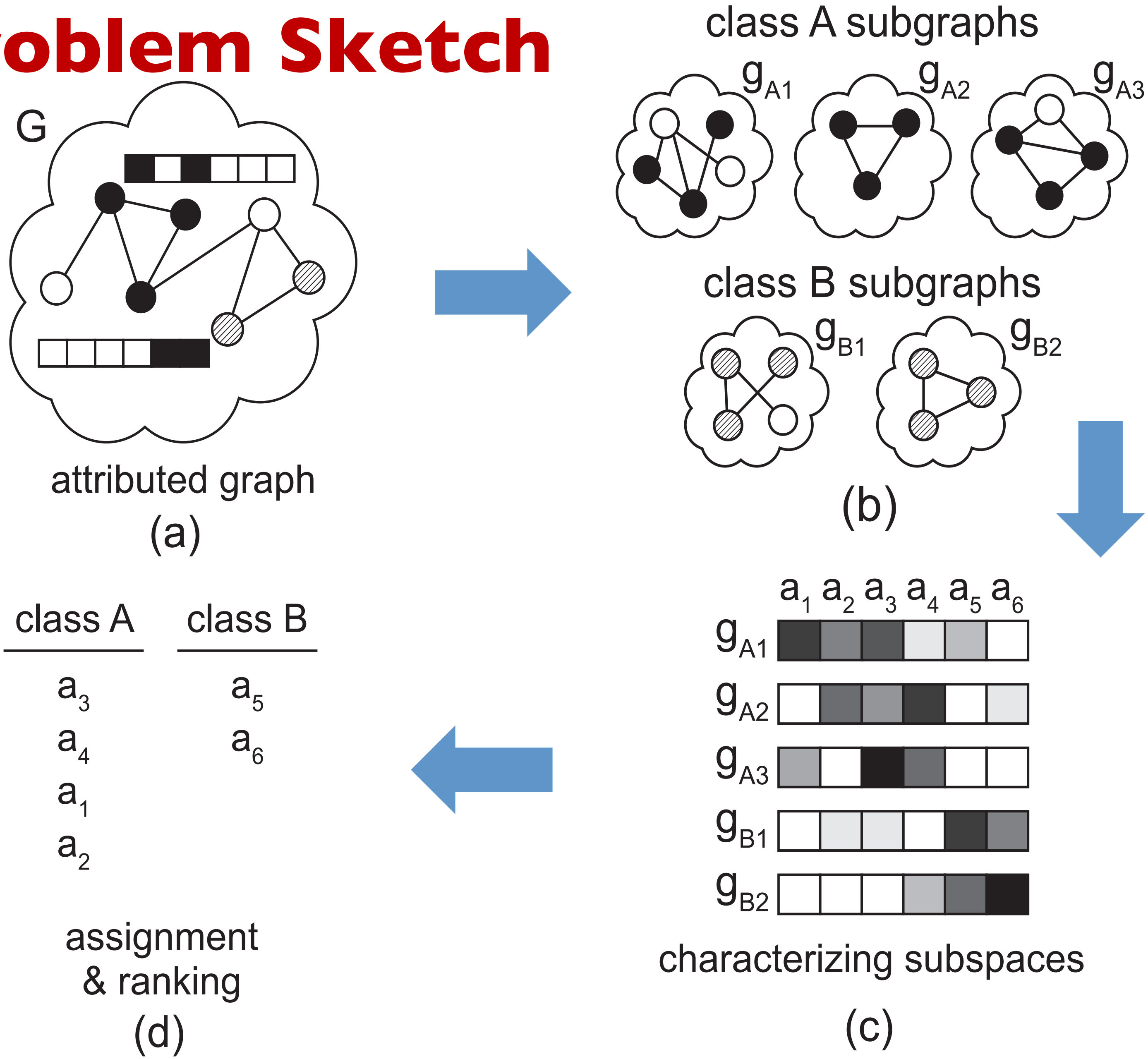
Crosswords
Tea parties
Gardening

Class A subgraphs

Class B subgraphs

Selfies
Partying
Video Games

Problem Sketch



Characterization Problem: Formal

Given

- p attributed subgraphs $g_1^+, g_2^+, \ldots, g_p^+$ from class 1, $\mathcal{S}^+$
- n attributed subgraphs $g_1^-, g_2^-, \ldots, g_n^-$ from class 2, $\mathcal{S}^-$ from graph G, and attribute vector $\mathbf{a} \in \Re^d$ for each node;

Find

- a partitioning of attributes to classes as A^+ and A^-, where $A^+ \cup A^- = A$ and $A^+ \cap A^- = \emptyset$,
- focus attributes $A_i^+ \subseteq A^+$ (and respective weights $\mathbf{w}_i^+$) for each subgraph g_i^+, $\forall i$, and
- focus attributes $A_j^- \subseteq A^-$ (and respective weights $\mathbf{w}_j^-$) for each subgraph g_j^-, $\forall j$;

such that

- total quality Q of all subgraphs is maximized, where $Q = \sum_{i=1}^{p} q(g_i^+|A^+) + \sum_{j=1}^{n} q(g_j^-|A^-)$;

Rank attributes within A^+ and A^-.

Reminder: Normality

- Normality as subgraph quality q:

$$N = w_c^T \cdot (\widehat{x_I} + \widehat{x_X})$$

$$\max_{w_c} \mathrm{N} \quad \text{s.t. } \left\|w_c\right\|_p = 1, w_c(a) \geq 0, \forall a = 1, \ldots, d$$

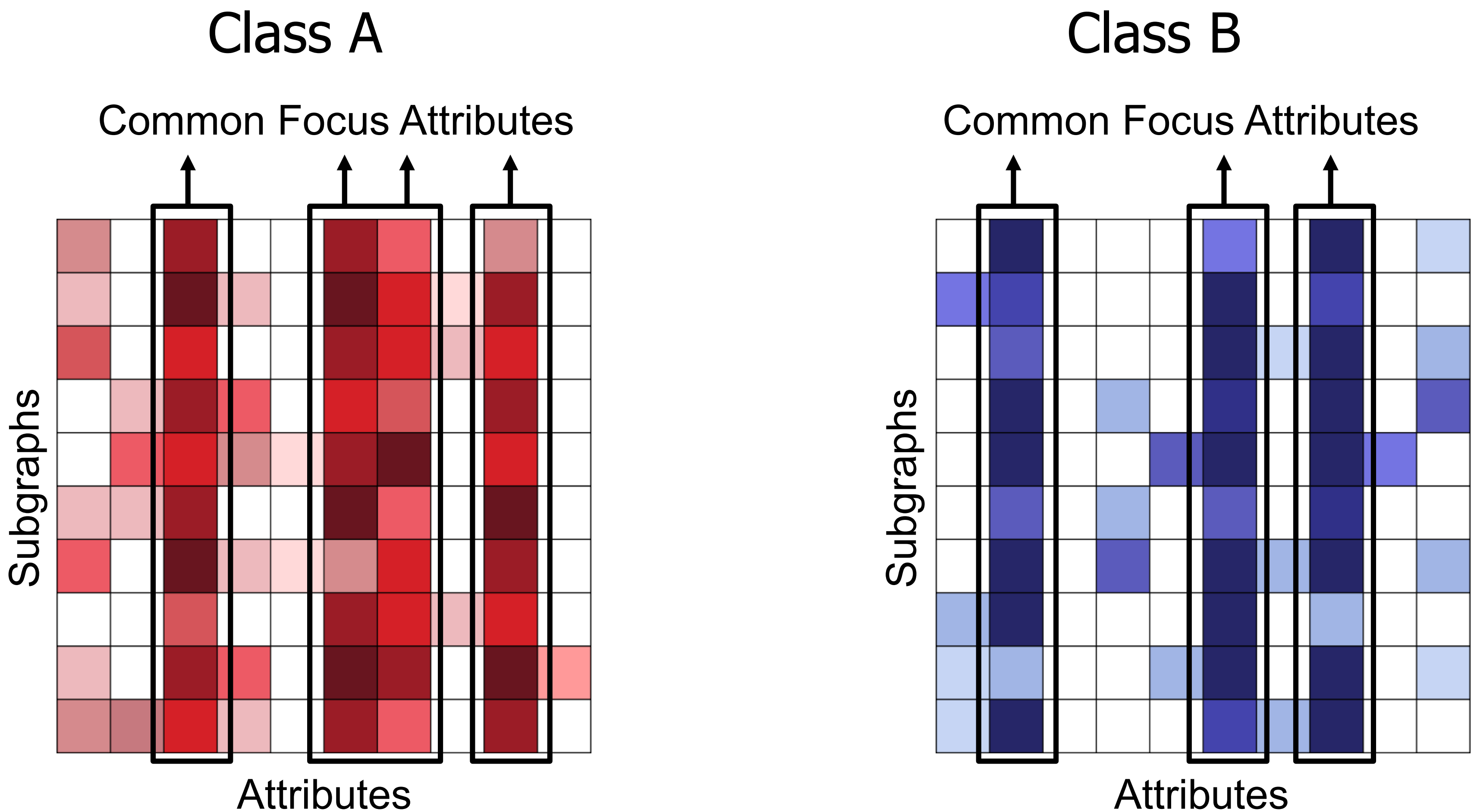
L_1 norm

$w_c(a) = 1$, **one** attribute with largest **x**

L_2 norm

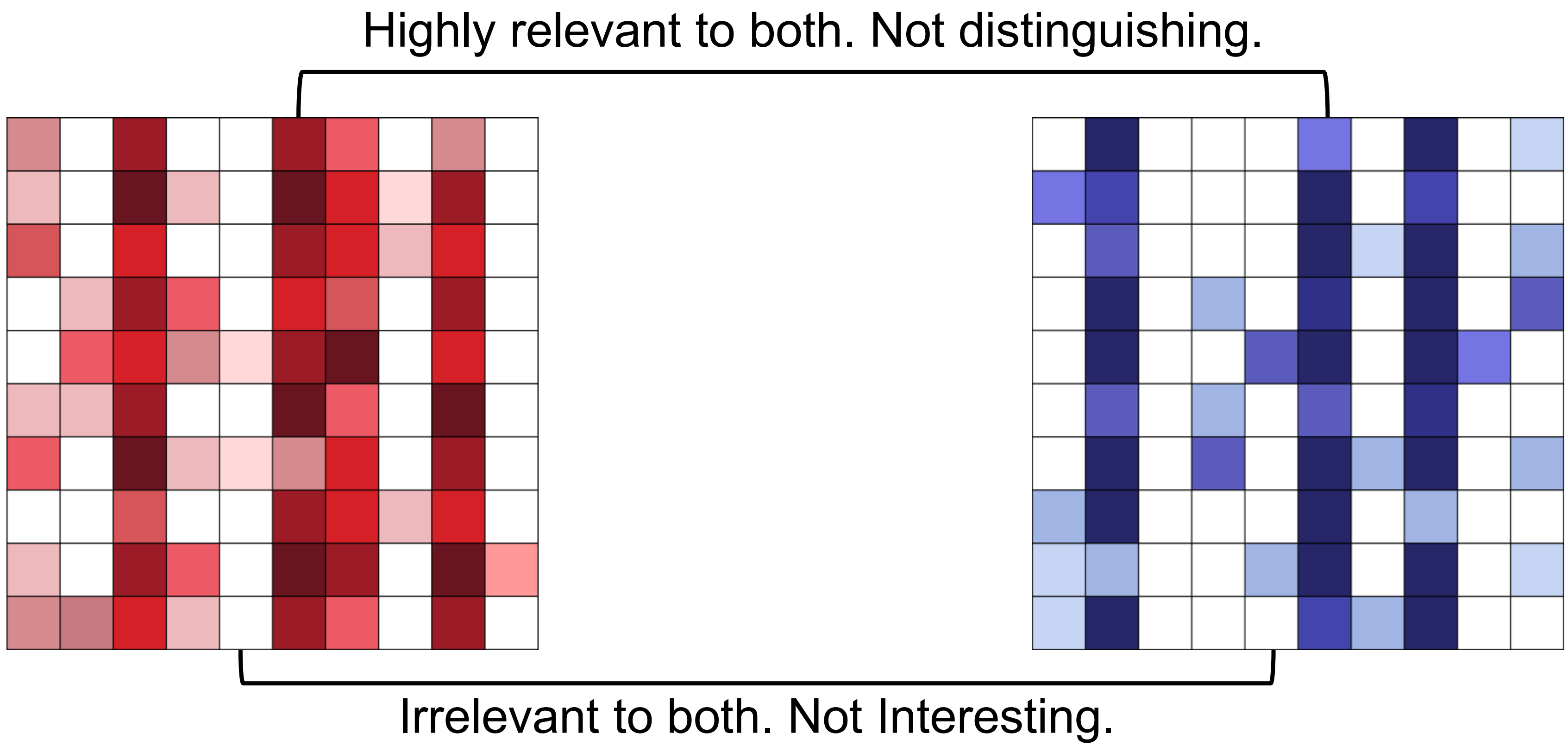
$w_c(a) = \frac{x(a)}{\sqrt{\sum_{x(i)>0} x(i)^2}}$, **all** attributes with positive **x**

Splitting attributes by class: intuition



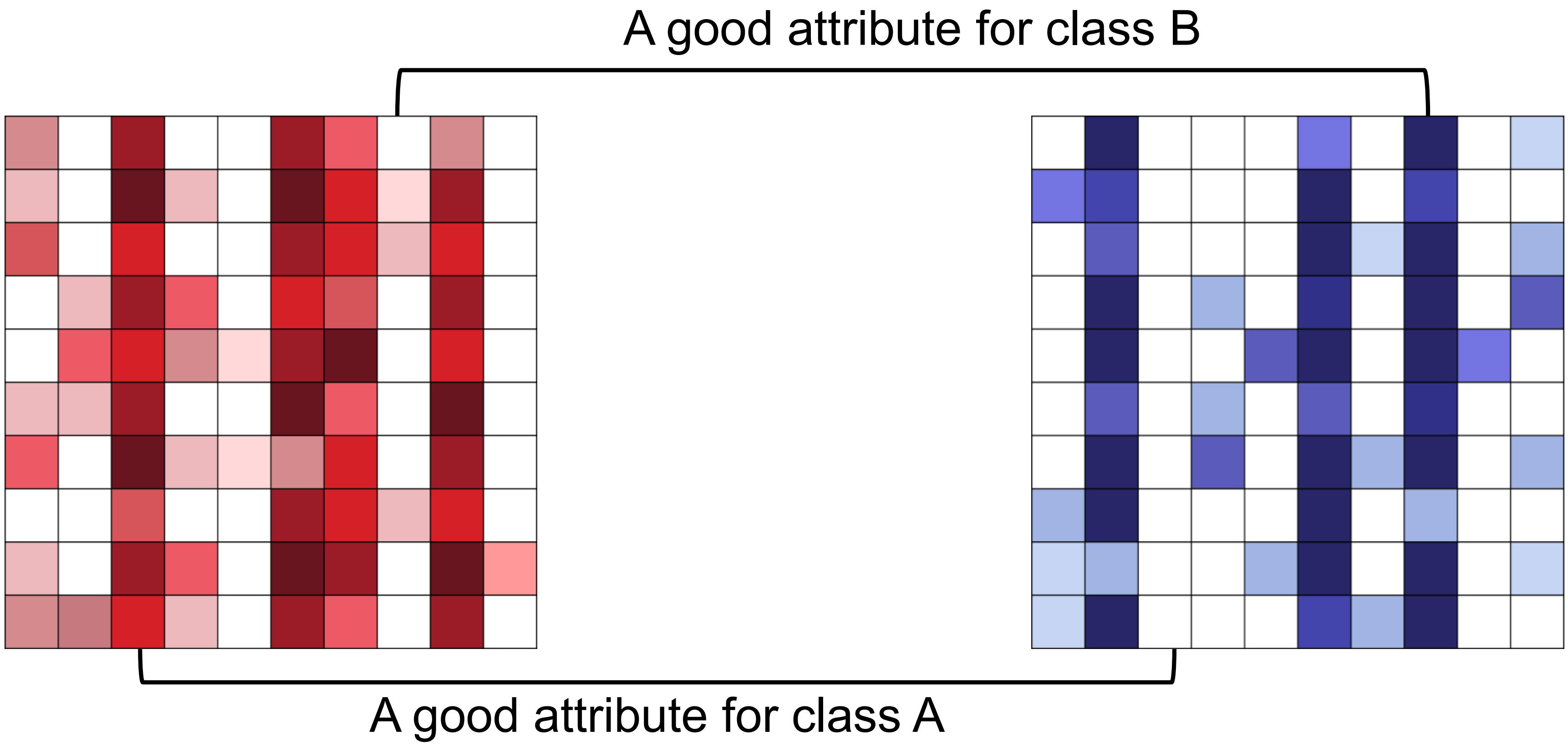
Splitting attributes by class: intuition

- We don't want attributes that are:
  - Relevant or irrelevant to ***both*** classes



Splitting attributes by class: intuition

- We want attributes that are:
  - Relevant to **one** class & irrelevant to other(s)



Setting up the objective

Details

- Given a subset of attributes *S*, normality of subgraph *g* is

$$\underbrace{N(g|S)}_{\substack{\text{2-norm of } x \text{ induced} \\ \text{on the attribute subspace}}} = \sqrt{\sum_{a \in S} x(a)^2} = \left\| x[S] \right\|_2$$

attribute weight vector of g

Setting up the objective

Details

- **Quality of an attribute split** is:

$$\max_{A^+ \subseteq A, A^- \subseteq A} \frac{1}{p} \sum_{i \in S^+} \left\| x_i[A^+] \right\|_2 + \frac{1}{n} \sum_{j \in S^-} \left\| x_j[A^-] \right\|_2$$

Such that $A^+ \cap A^- = \emptyset$

p = number of subgraphs in class +
n = number of subgraphs in class -

Setting up the objective

Details

- **Quality of an attribute split** is:

$$\max_{A^+ \subseteq A, A^- \subseteq A} \frac{1}{p} \sum_{i \in S^+} \left|\left| x_i[A^+] \right|\right|_2 + \frac{1}{n} \sum_{j \in S^-} \left|\left| x_j[A^-] \right|\right|_2$$

Such that $A^+ \cap A^- = \emptyset$

- **Rank attributes** by

p = number of subgraphs in class +
n = number of subgraphs in class -

$$rc(a) = \underbrace{\frac{1}{p} \sum_{i \in S^+} x_i(a)}_{\text{Normalized contribution of } a \text{ to Class +}} - \underbrace{\frac{1}{n} \sum_{j \in S^-} x_j(a)}_{\text{Normalized contribution of } a \text{ to Class -}}$$

Submodular Welfare Problem

Details

- Definition:

*Given **d** items and **m** players having a **monotone** and **submodular** utility function (w_i) over subsets of items. Partition the d items into m **disjoint sets** $(I_1, I_2, \cdots, I_m)$ in order to maximize:*

$$\sum_{i=1}^{m} w_i(I_i)$$

- Our quality function $N(g|S)$ is a ***monotone*** and ***submodular*** set function.

$$w_c(I_c) = N(\mathcal{S}^{(c)}|A^{(c)}) = \frac{1}{n^{(c)}} \sum_{k \in \mathcal{S}^{(c)}} \|\mathbf{x}_k[A^{(c)}]\|_2$$

Attribute splitting as SWP

Details

- SWP is **NP-hard**
- First approx. factor is ½ [Lehmann+, 2001]
- Improved to $(1 - 1/e)$ [Vondrák+, 2008]
- No better approximation unless
  - P = NP [Khot+, 2008]
  - Using **exponentially-many** value queries [Mirrokni+, 2008]

→ [Vondrák+, 2008] is optimal approximation

Experiments

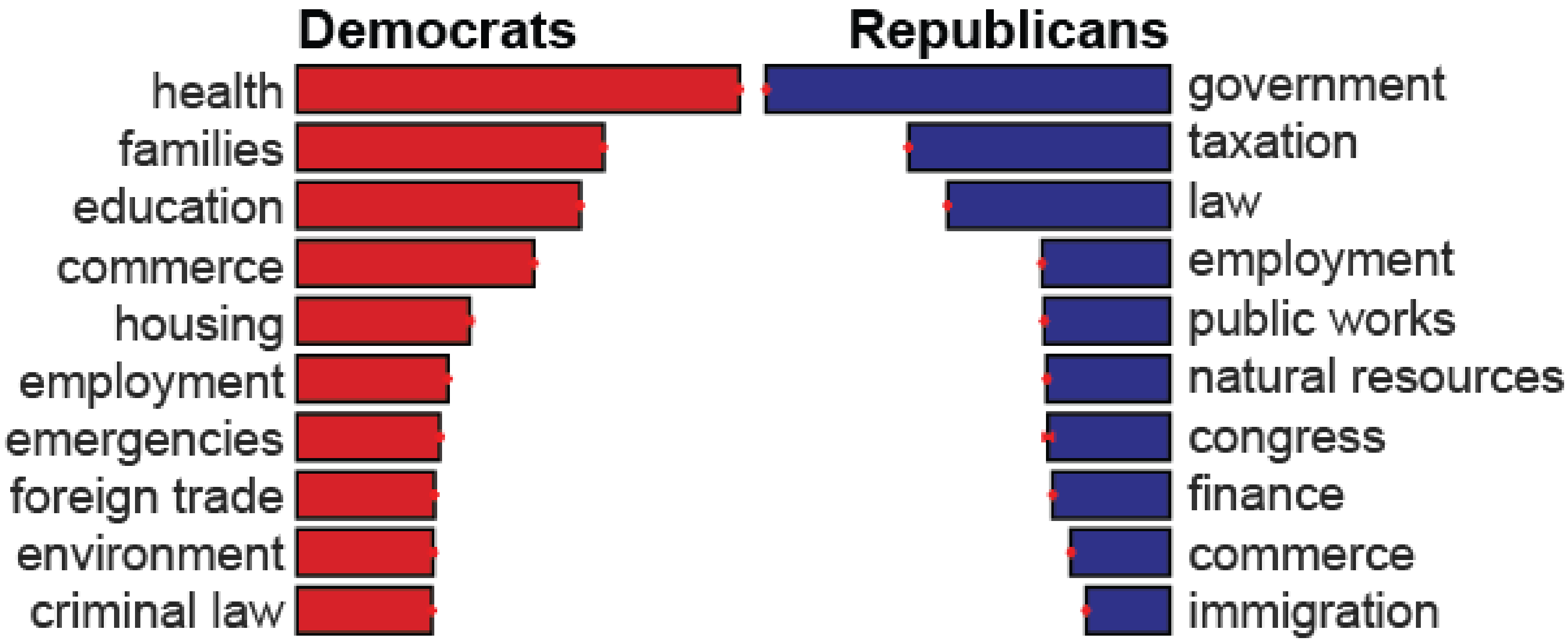
- Datasets
  - Congress Co-sponsorship Network
  - Amazon Co-purchase Network
  - DBLP Co-authorship Network

- Baseline (LASSO): L1-Regularized Logistic Regression
  - Positive weights are assigned to class **A**
  - Negative weights are assigned to class **B**

Congress Co-sponsorship

- Bills in Congress
  - each bill has a set of ***sponsors*** & ***policy area tag***
- **Attributed Graph:**
  - Nodes: **congressmen**
  - Edges: ***co-sponsoring*** a bill
  - Attributes: ***policy areas*** of bills they sponsored:
    - National Security and Armed Forces
    - Environmental Protection
    - Foreign Affairs
    - …
- Classes: ***party affiliation*** of congressmen

Liberal and Conservative Ideals

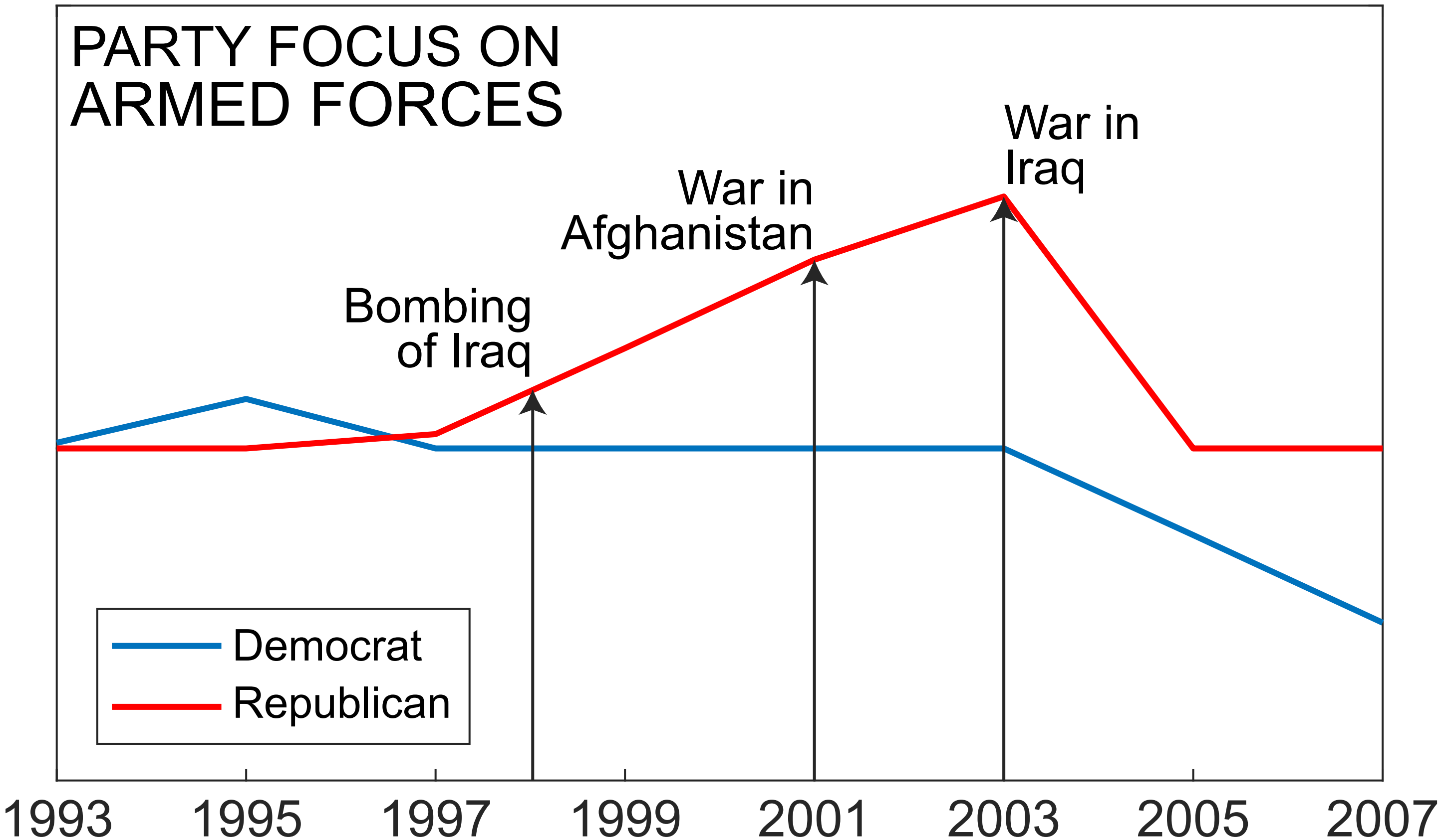


Democrats focus mostly on ***social*** programs

Republicans focus mostly on ***governance*** and ***finance***

Focus Over Time

- 13 consecutive congress two-year cycles:

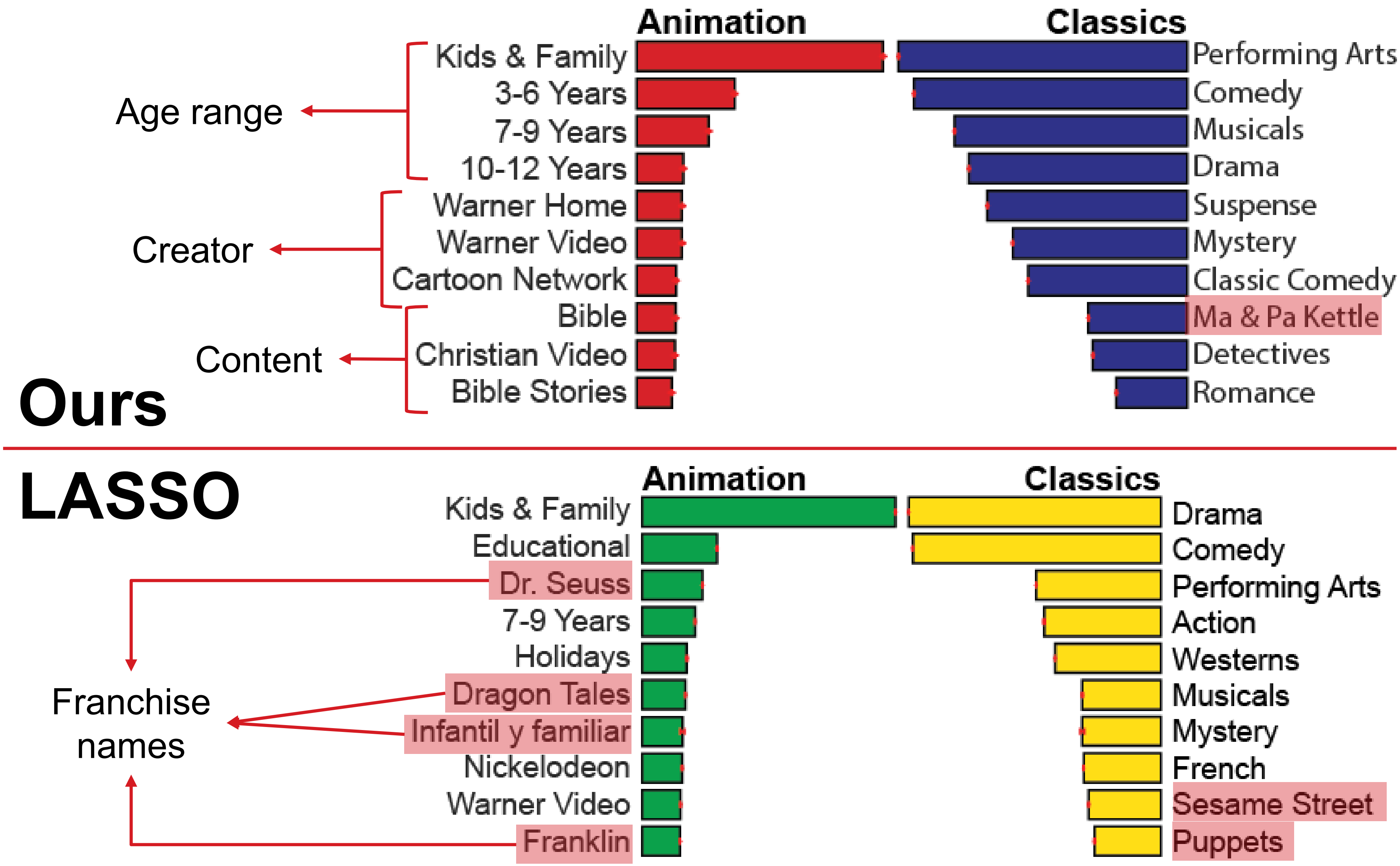


Amazon.com Co-purchases

Attributed Graph:

- Nodes: Amazon **videos**
- Edges: being **co-purchased** together
- Attributes:
  - Product genre (Drama, Comedy, etc.)
  - Audience age range (e.g., 10-12 years)
  - Creators (e.g. Warner Video)
  - …

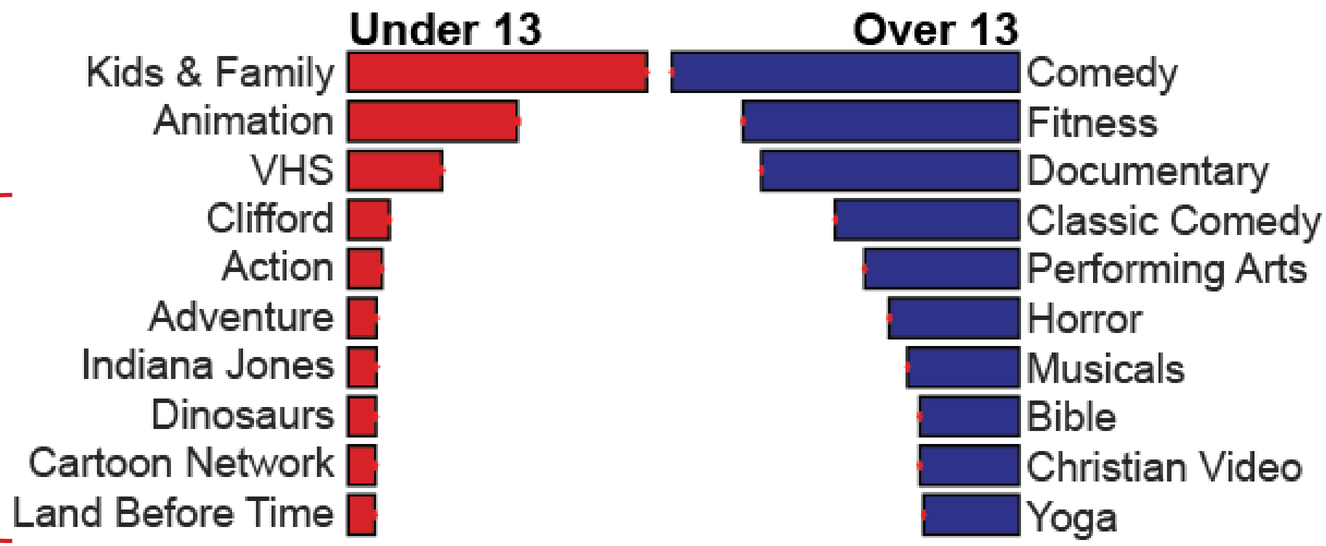
Classes: Animation vs. Classic



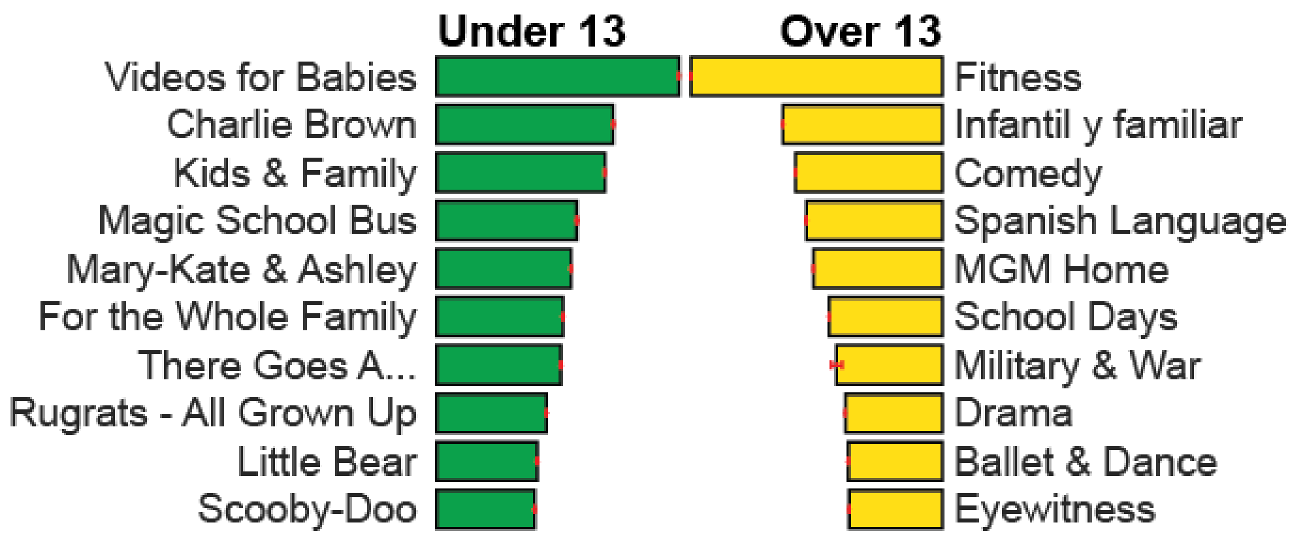
Classes: Under 13 vs. Over 13

Attribute weight goes down as quality decreases

Ours



LASSO



Not much differentiation

Characterization vs. Classification

- Regularized linear classifiers (e.g. LASSO) can find
  - a sparse attribute subspace
  - coefficients for ranking
- How is our work different?

Classifiers focus on ***confidence***
while we focus on ***support***

Characterization vs. Classification

Confidence → Prob. of belonging to class c if a is observed

$$Cfd(c,a) = \Pr(c|a) = \frac{\#(c,a)}{\#(a)}$$

Support → Portion of nodes in class c exhibiting a

$$Sup(c,a) = \frac{\#(c,a)}{\#(c)}$$

Characterization vs. Classification

Class Confidence → *Relative* Confidence

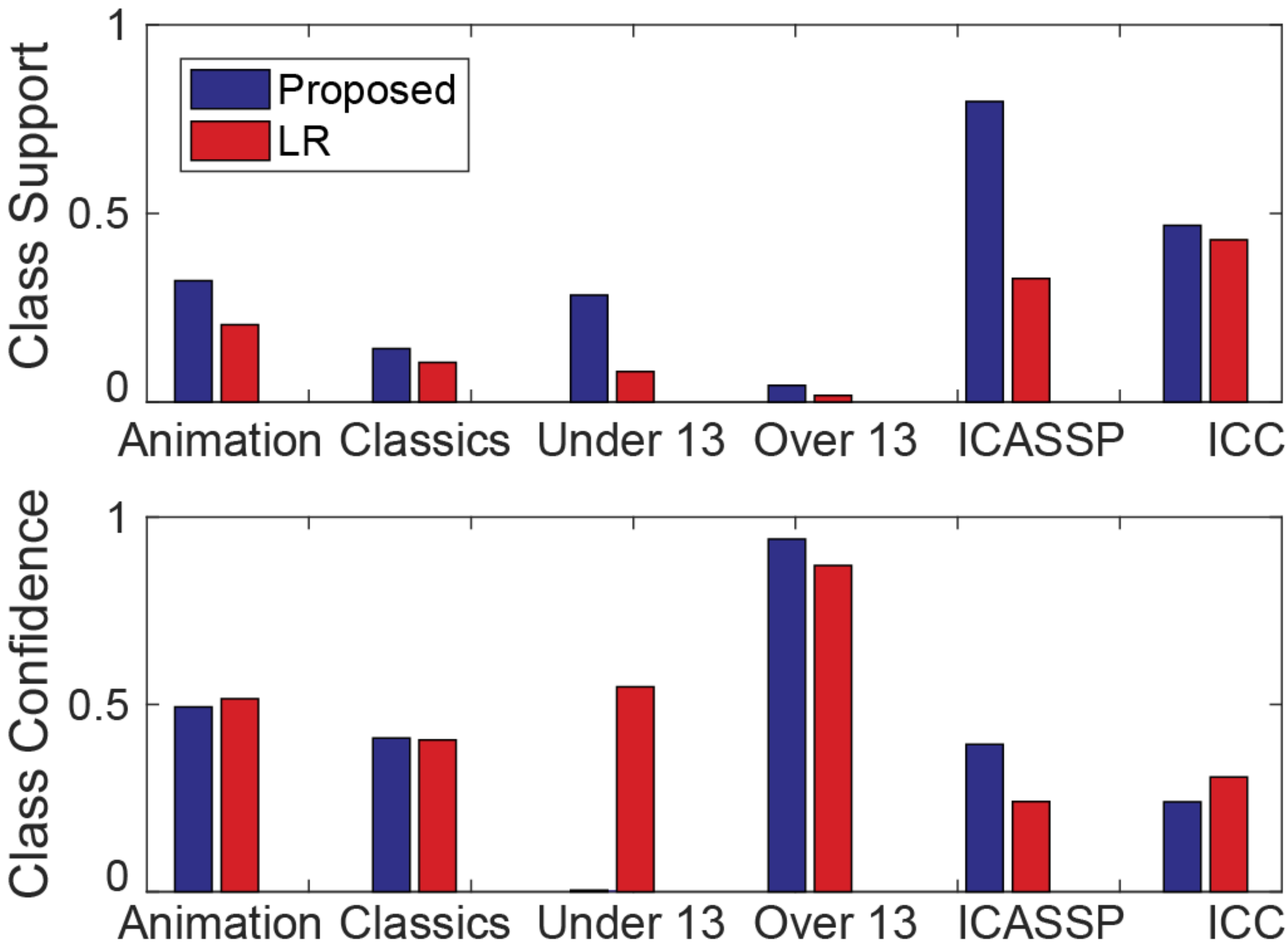
$$CC(c^+, a) = \Pr(c^+|a) - \Pr(c^-|a)$$

Class Support → *Relative* Support

$$CS(c^+, a) = Sup(c^+, a) - Sup(c^-, a)$$

Classifiers focus on ***confidence***
while we focus on ***support***

Characterization vs. Classification

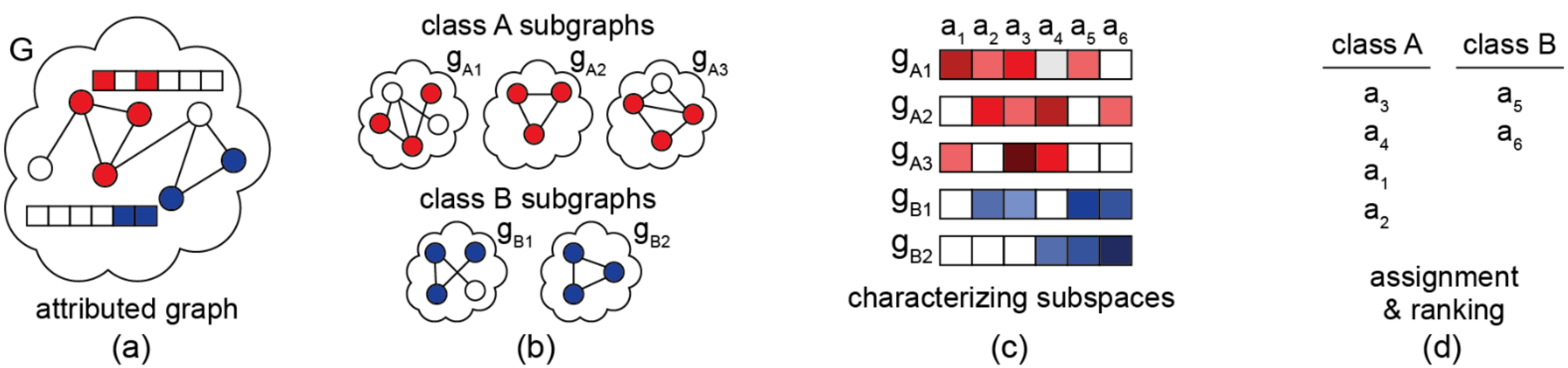


Slides, code, data http://www3.cs.stonybrook.edu/~arezaei/project/amen_char.html

Characterizing Class Differences in Attributed Graphs

Aria Rezaei, Bryan Perozzi, Leman Akoglu

Overview

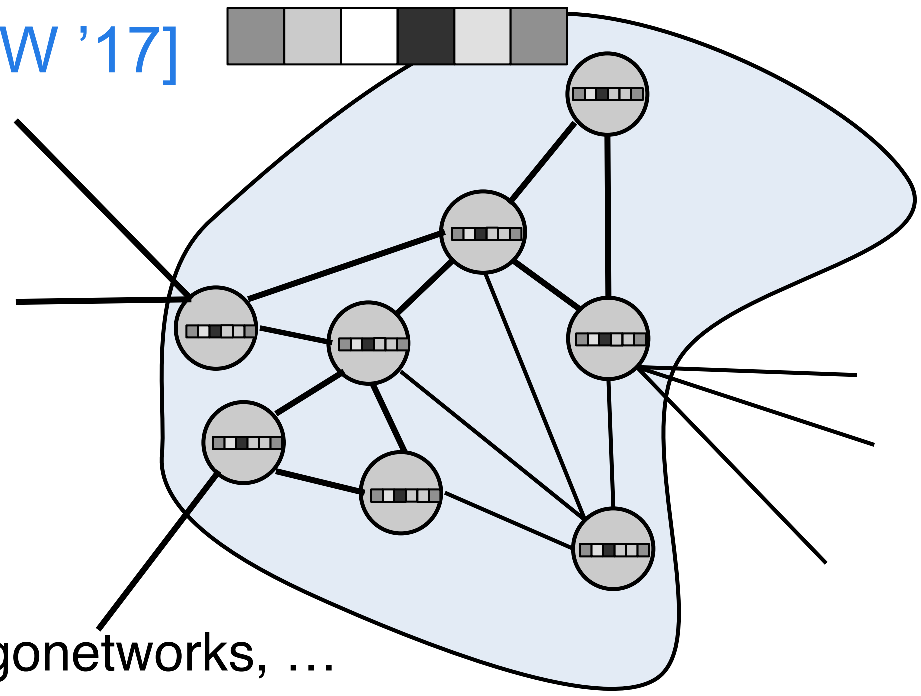


Ties That Bind - Characterizing Classes by Attributes and Social Ties. *Aria Rezaei, Bryan Perozzi, Leman Akoglu.* WWW 2017 Companion

Slides

This talk

- Attributed (sub)graphs*
  - Subgraphs [SIAM SDM'16]
  - Summarization [ACM TKDD'18]
  - Comparisons [WWW '17]



* social circles, communities, egonetworks, …

References, Links to Code&Data:

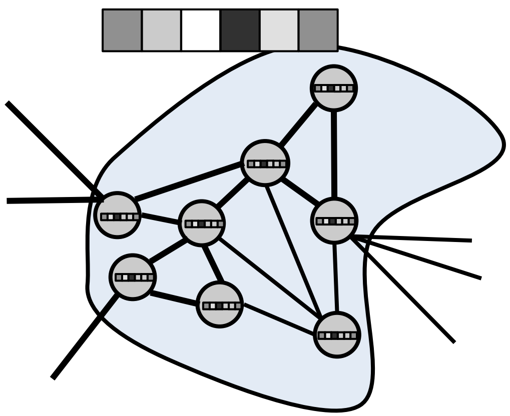
- **Scalable Anomaly Ranking of Attributed Neighborhoods.** *Bryan Perozzi and Leman Akoglu.* SIAM SDM 2016
  https://github.com/phanein/amen

- **Discovering Communities and Anomalies in Attributed Graphs: Interactive Visual Exploration and Summarization.** *Bryan Perozzi and Leman Akoglu.* ACM TKDD, 2018
  https://www.dropbox.com/home/Public/iSCAN

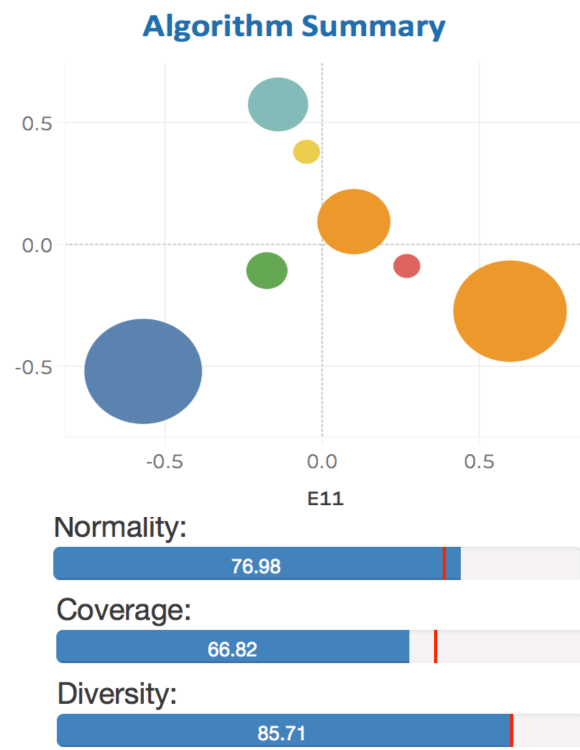
- **Ties That Bind - Characterizing Classes by Attributes and Social Ties.** *Aria Rezaei, Bryan Perozzi, Leman Akoglu.* WWW 2017 Companion
  https://github.com/rezaeiaria/AmenChar

Contact: lakoglu@andrew.cmu.edu
www.andrew.cmu.edu/~lakoglu

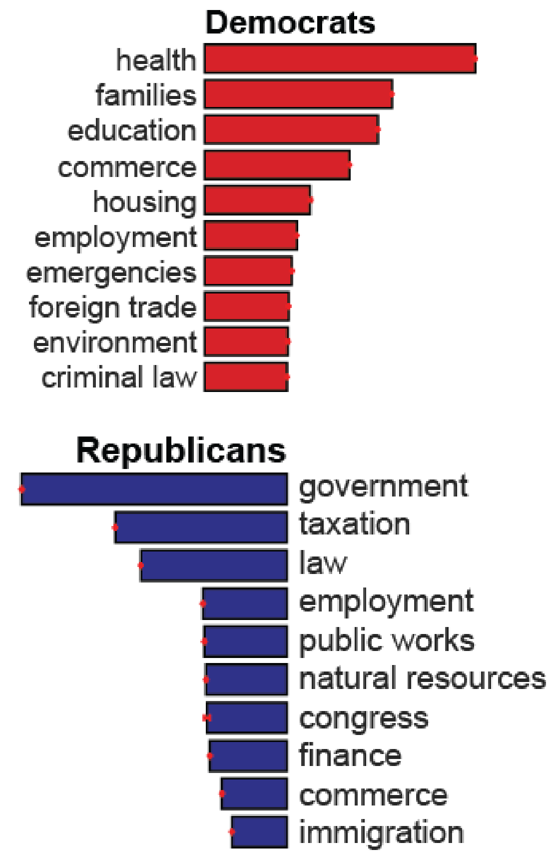
Subgraphs



Summarization



Comparisons



Thanks!