Understanding Node-Attributed Networks: Interactive Exploration & Summarization

Leman Akoglu

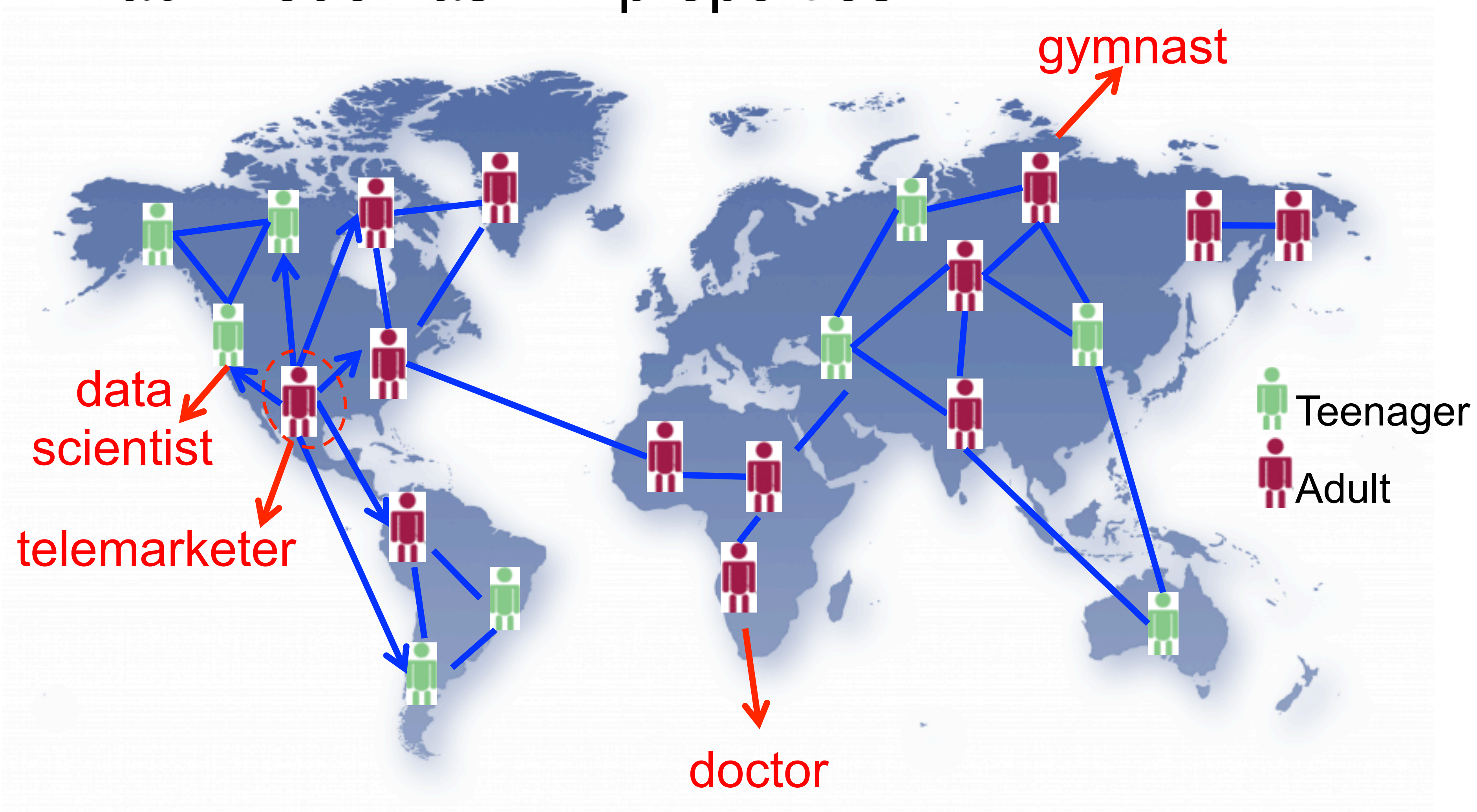
Joint work with Bryan Perozzi (Google Research NYC),
Rashmi Raghunandan, Shruti Sridhar, Upasna Suman (CMU)

Interactive Data Exploration and Analytics (IDEA)
August 14, 2017

Carnegie Mellon University
Heinz College

Attributed networks

Each node has 1+ properties



Attributed networks

Edge view

Source ID	Destination ID
4	1
44	1
195	1
197	1
9	2
15	2
20	2
30	2
.	.
.	.
.	.
.	.

Node view

ID	Age	Dorm	Major	.	.
2	23	130	10	.	.
3	26	134	. 23	.	.
7	25	133	34	.	.
5	22	140	45	.	.
.	.	.	.		
.	.	.	.		
.	.	.	.		
.	.	.	.		
.	.	.	.		

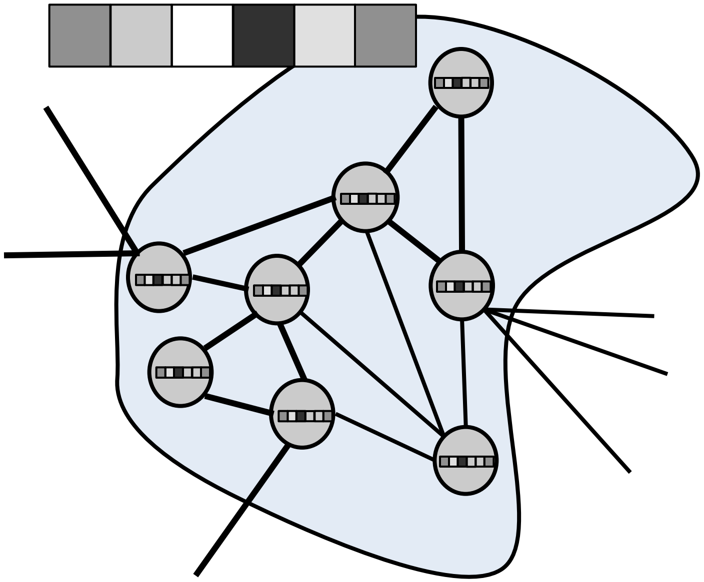
- Social networks demographics, likes, …
- Gene interactions ontological properties
- Web page properties
- …

Research question:

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

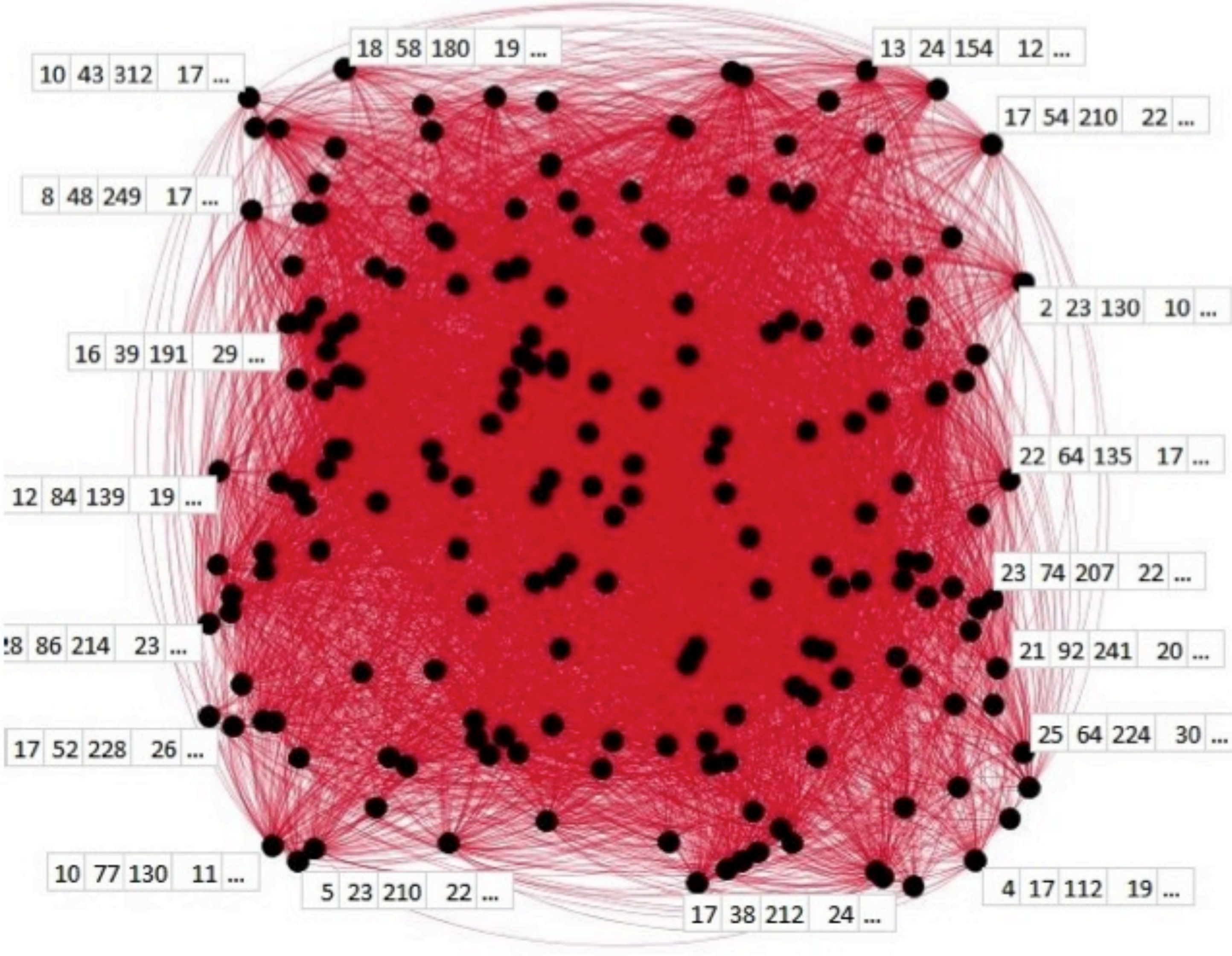
- describe
- characterize
- summarize

  <u>succintly</u>



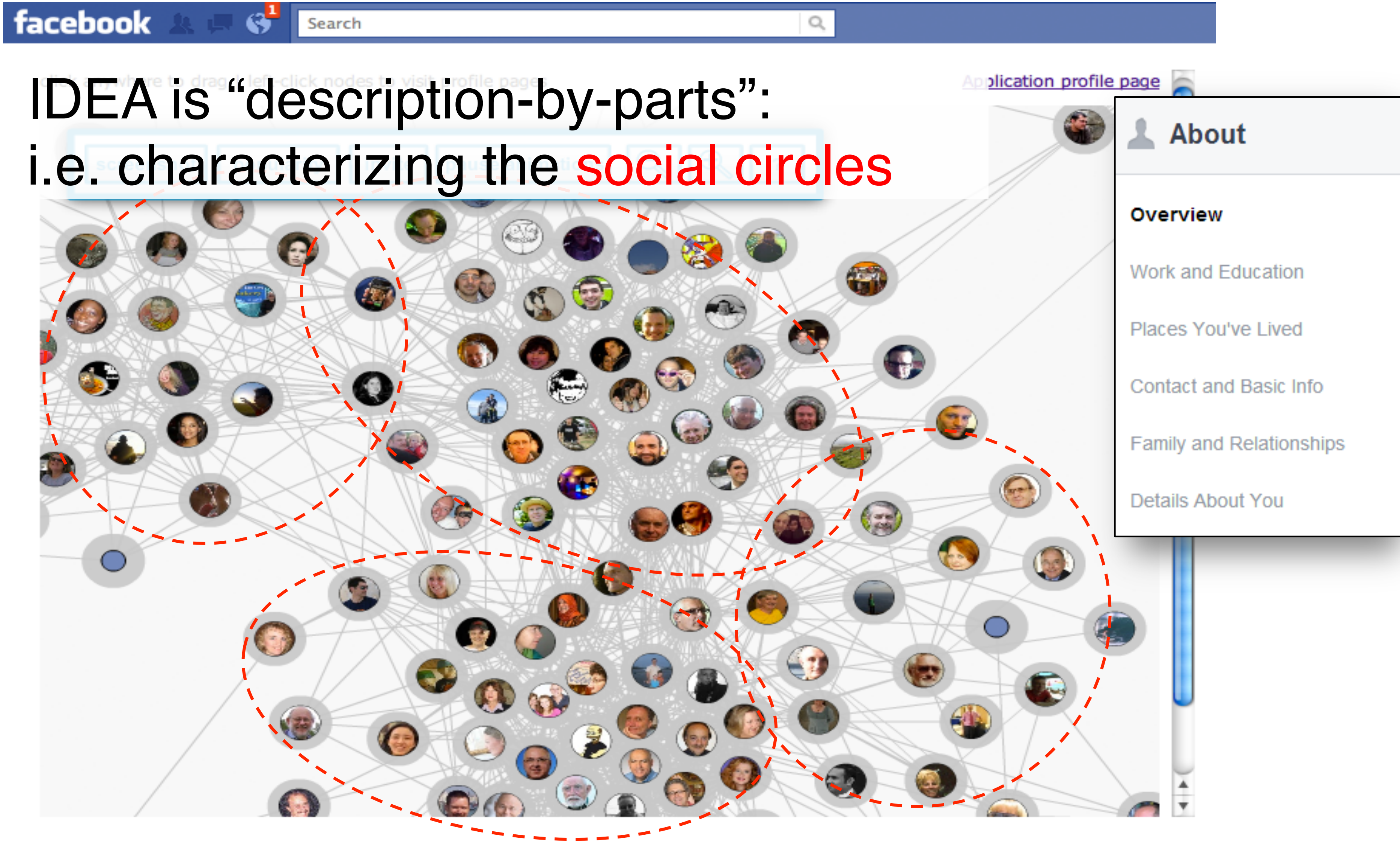
Attributed networks

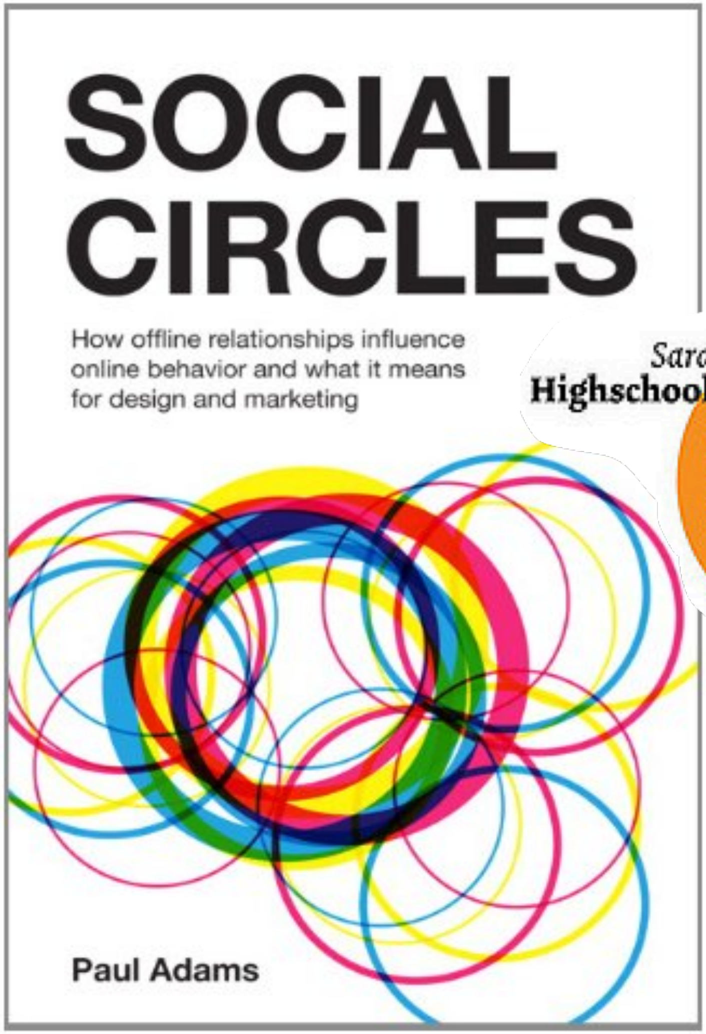
220 nodes, 6215 edges

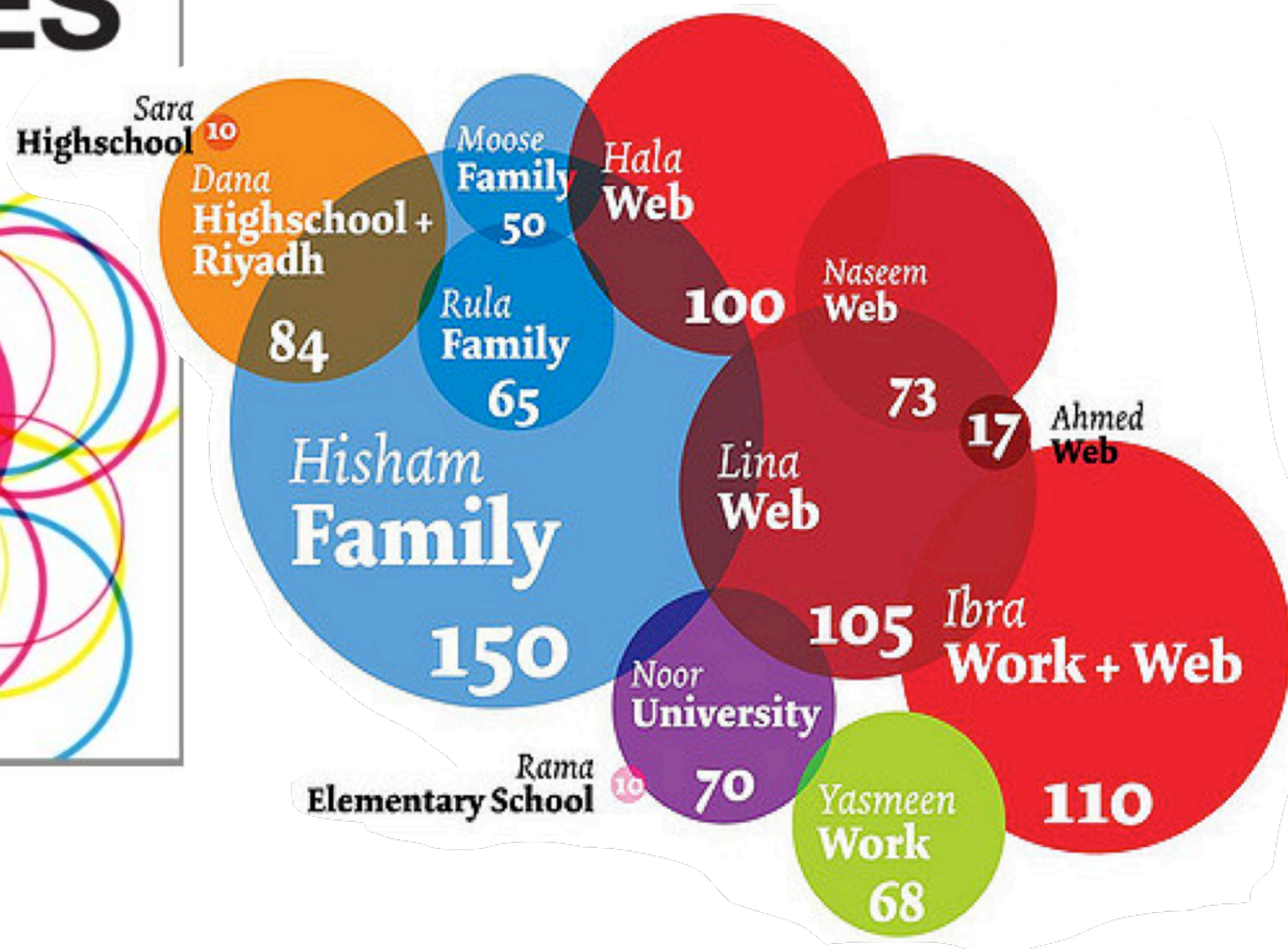


Attributed networks

IDEA is “description-by-parts”:
i.e. characterizing the social circles




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
Hala
Web
100
Naseem
Web
73
Ahmed
Web
17
Hisham
Family
150
Lina
Web
105
Ibra
Work + Web
110
Noor
University
70
Rama
Elementary School
10
Yasmeen
Work
68

Research question:

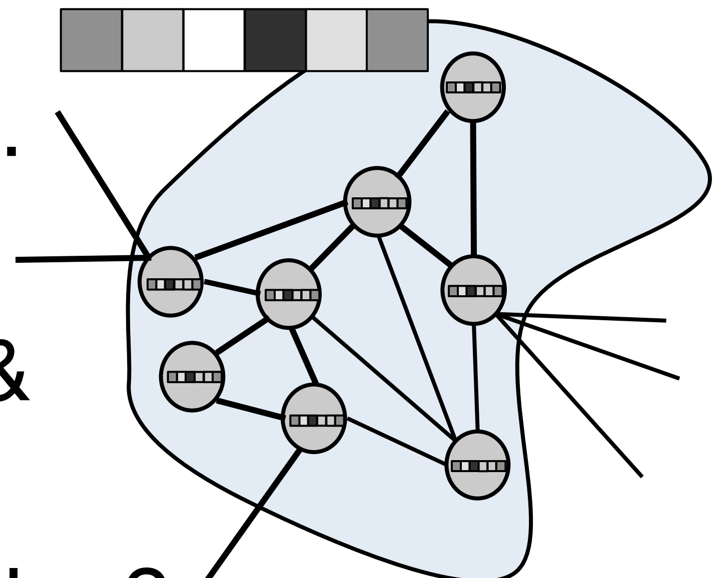
How can we **make sense** of node-attributed (social) networks **?**

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

② How to **extract** …

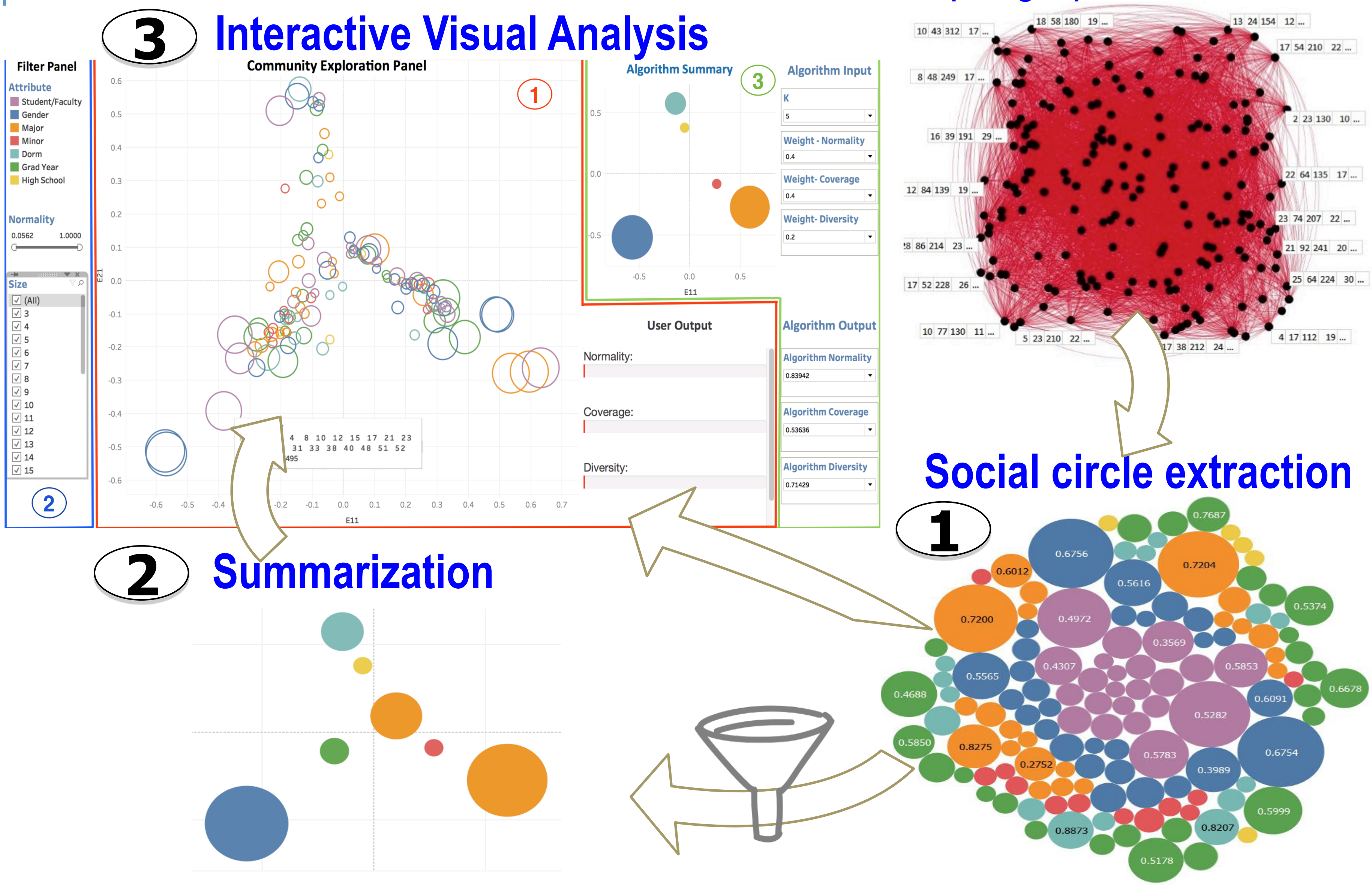
③ How to **visually explore** & **interactively summarize**

… social circles?

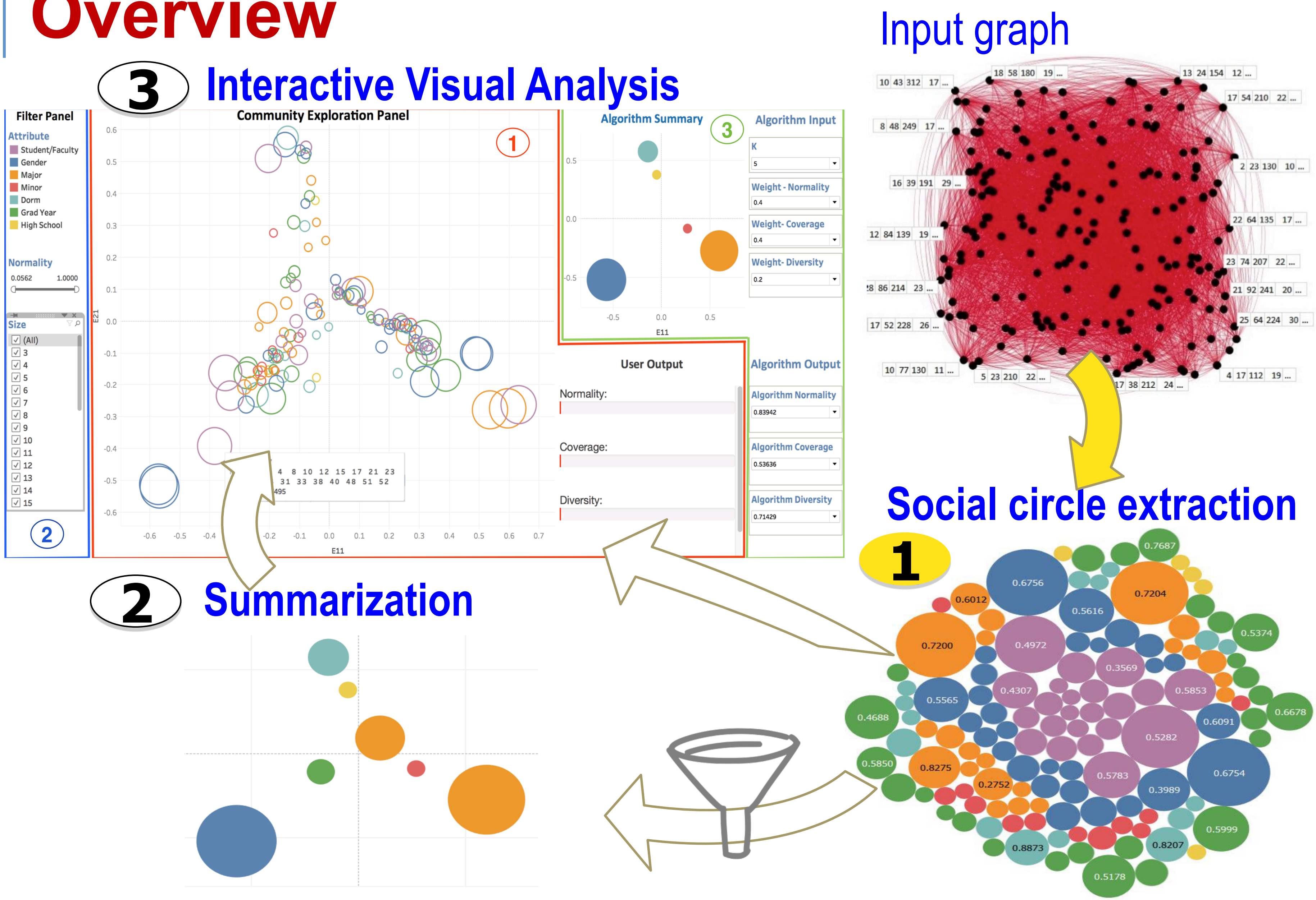


Overview

Input graph



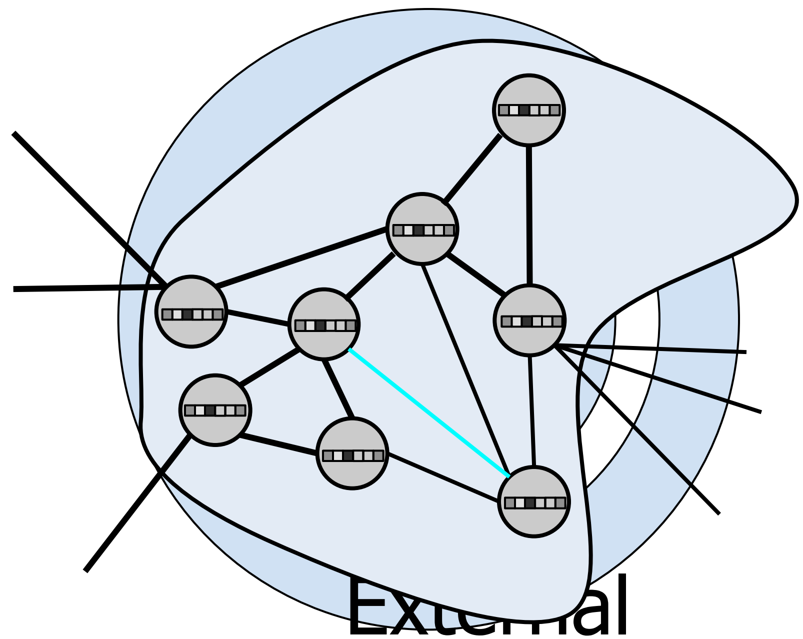
Overview



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
1
Social circle extraction
2
Summarization

What's a Social Circle, Anyhow?

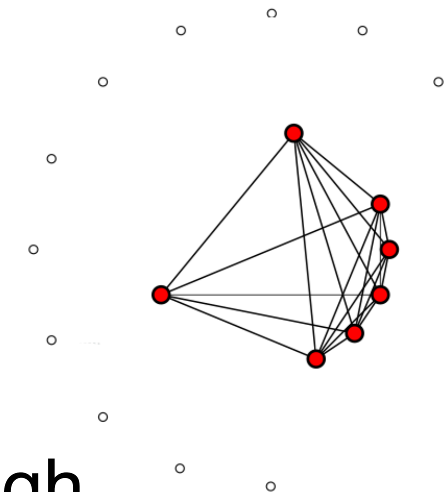
- **Given** an attributed subgraph, how to **quantify** its quality?
  - Structure-only
    - Internal-only
      - average degree
    - Boundary-only
      - cut edges
    - Internal + Boundary
      - conductance
  - Structure + Attributes



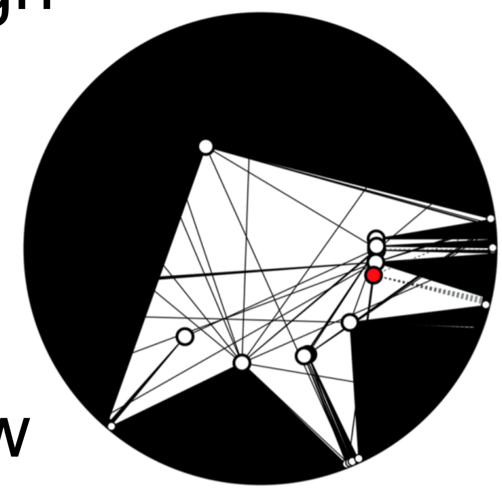
Scalable Anomaly Ranking of Attributed Neighborhoods
Bryan Perozzi and Leman Akoglu SIAM SDM 2016.

Normality (intuition)

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



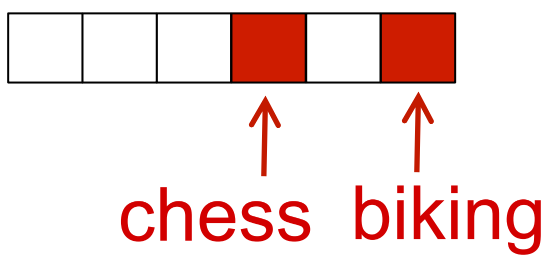
high

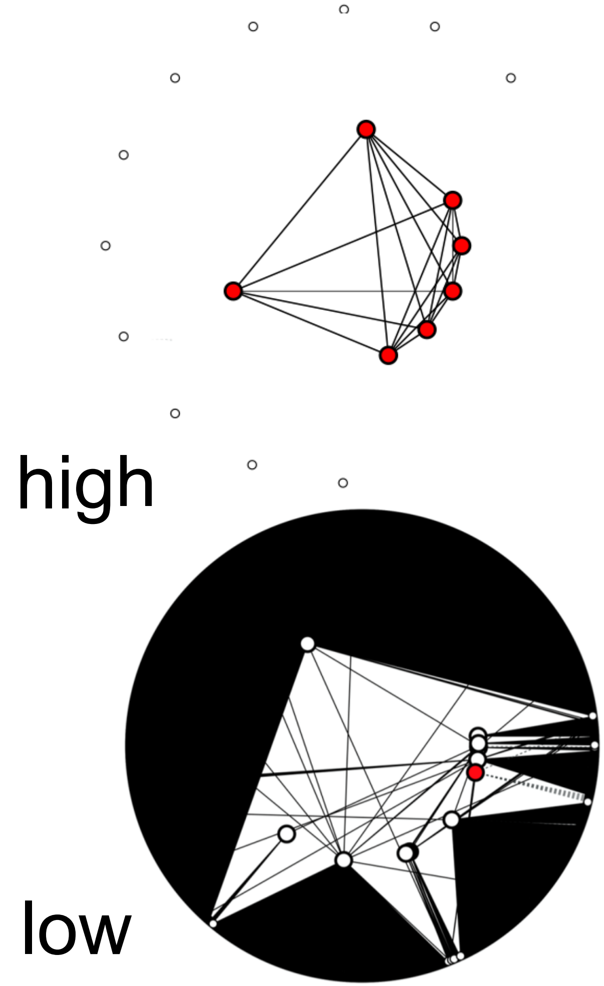


low

Normality (intuition)

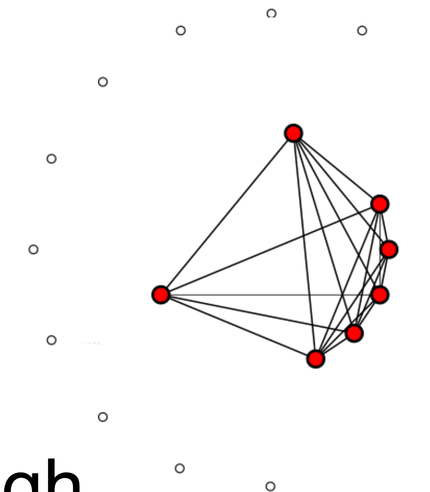
- **Given** an attributed subgraph
  how to **quantify** quality?
  - Internal
    - structural density AND
    - attribute coherence
      - *neighborhood **"focus"***



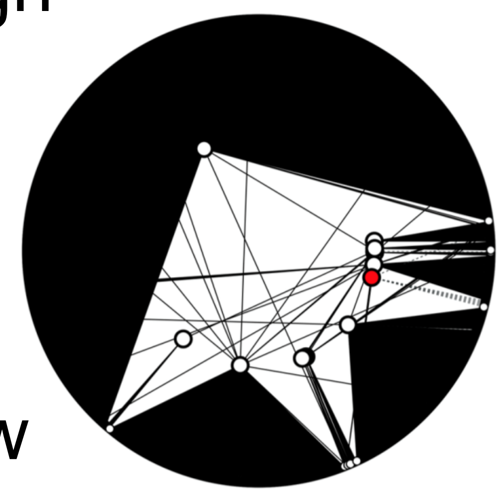


Normality (intuition)

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



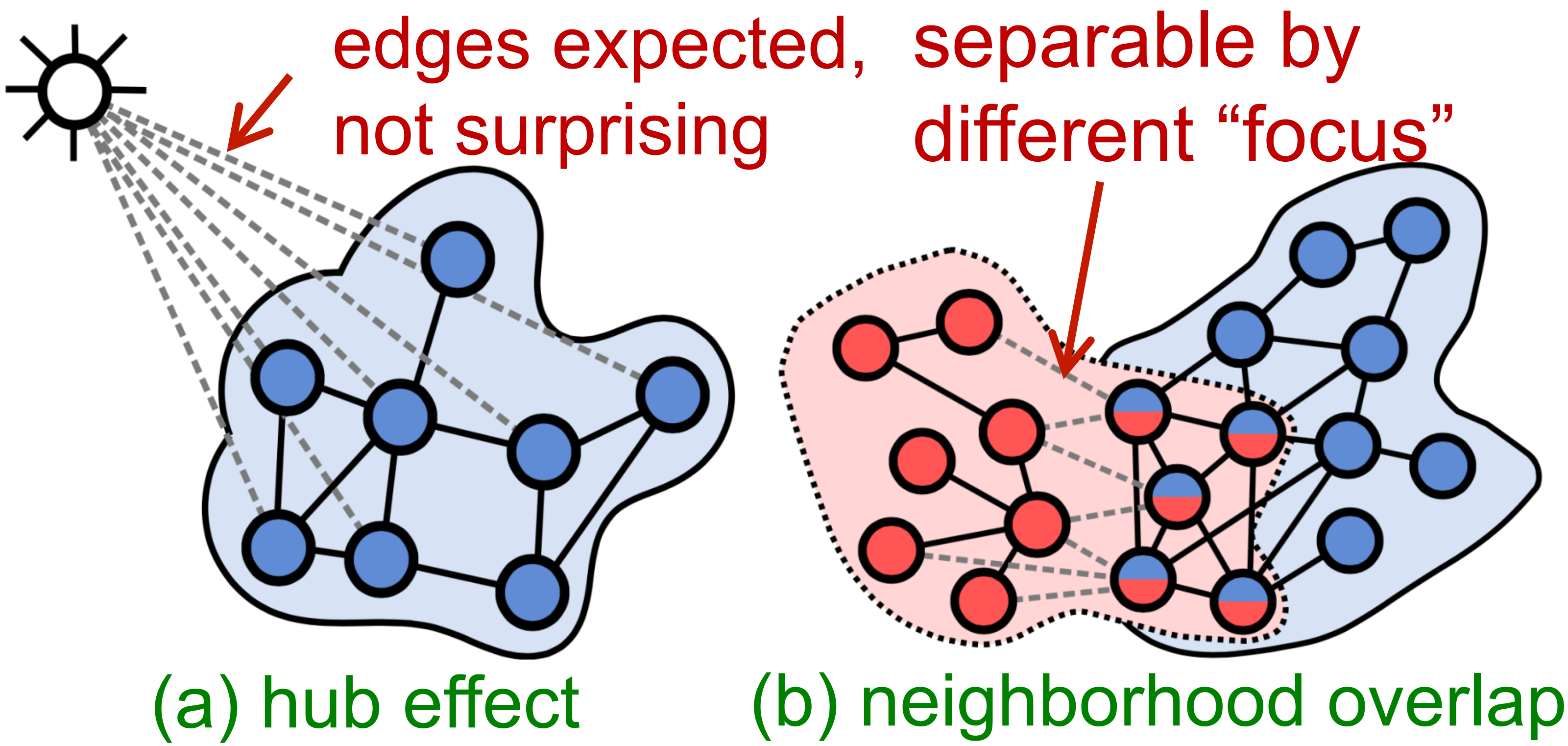
high



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

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

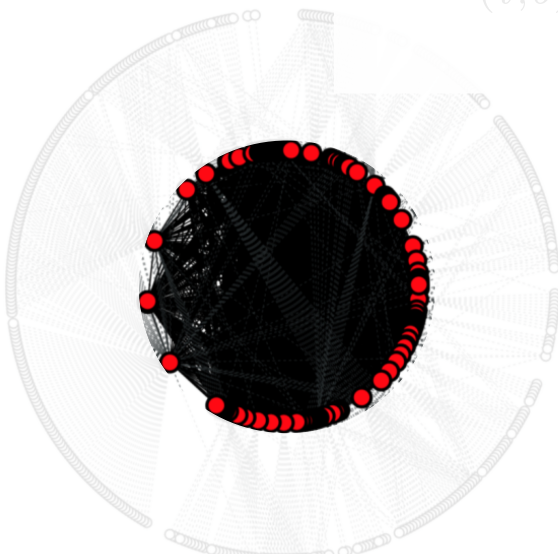
internal consistency

dot-product, or Kronecker's δ

"focus" vector

chess biking

1



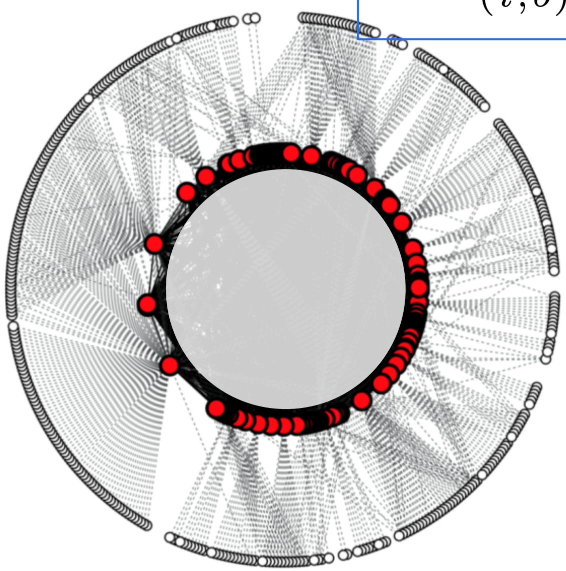
The measure of Normality

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

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

1

external separability



The measure of Normality

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

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

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

latent

Optimizing Normality

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

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

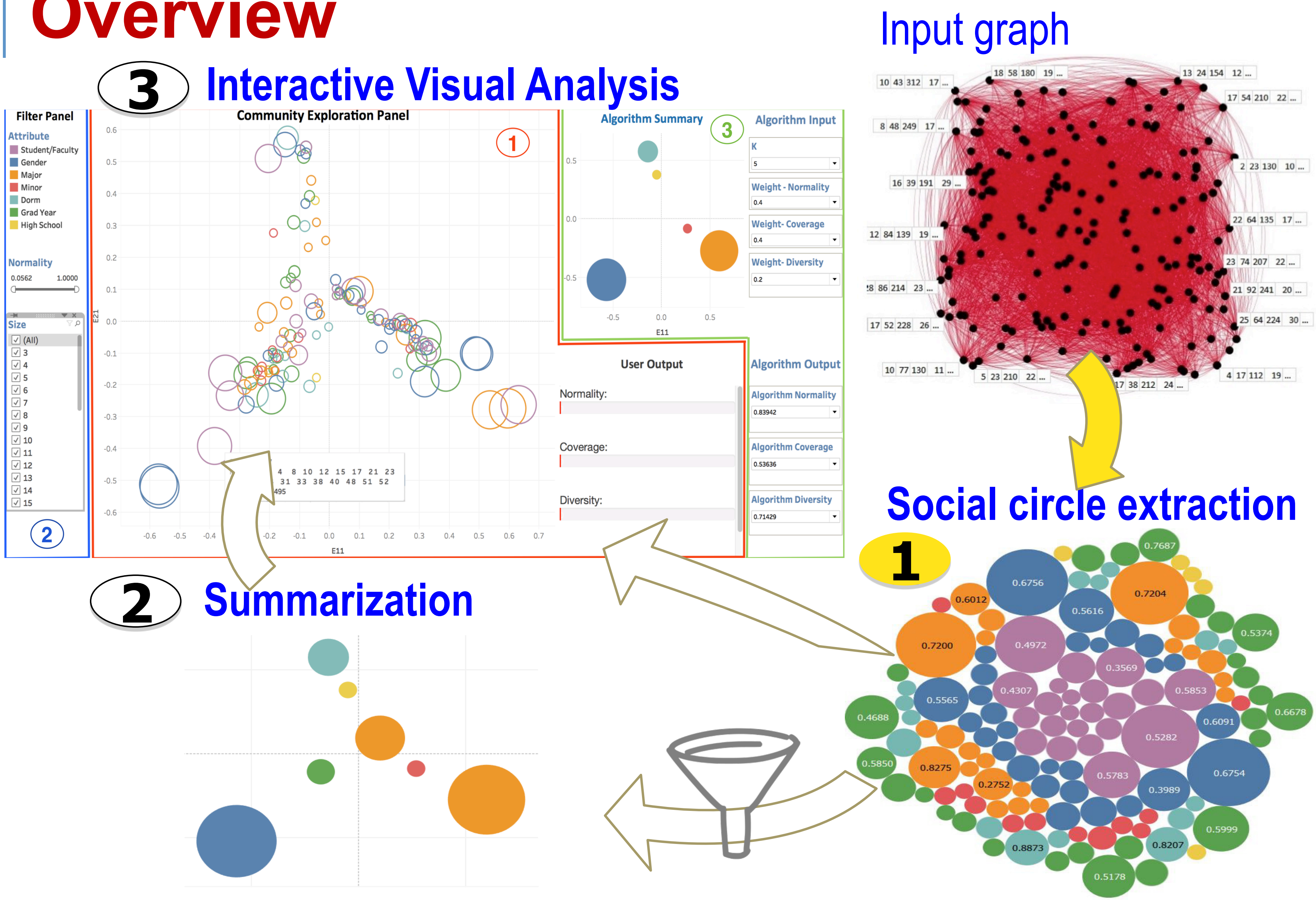
$p = 1$: $\mathbf{w_C}(f) \ = \ 1$ **one** attribute f with largest **x**

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

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

Linear in number of attributes!

Overview



Input graph
3
Interactive Visual Analysis
Filter Panel
Attribute
Student/Faculty
Gender
Major
Minor
Dorm
Grad Year
High School
Normality
Size
Community Exploration Panel
Algorithm Summary
Algorithm Input
K
Weight - Normality
Weight- Coverage
Weight- Diversity
User Output
Normality:
Coverage:
Diversity:
Algorithm Output
Algorithm Normality
Algorithm Coverage
Algorithm Diversity
1
2
Summarization
Social circle extraction

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

Extracting Social Circles

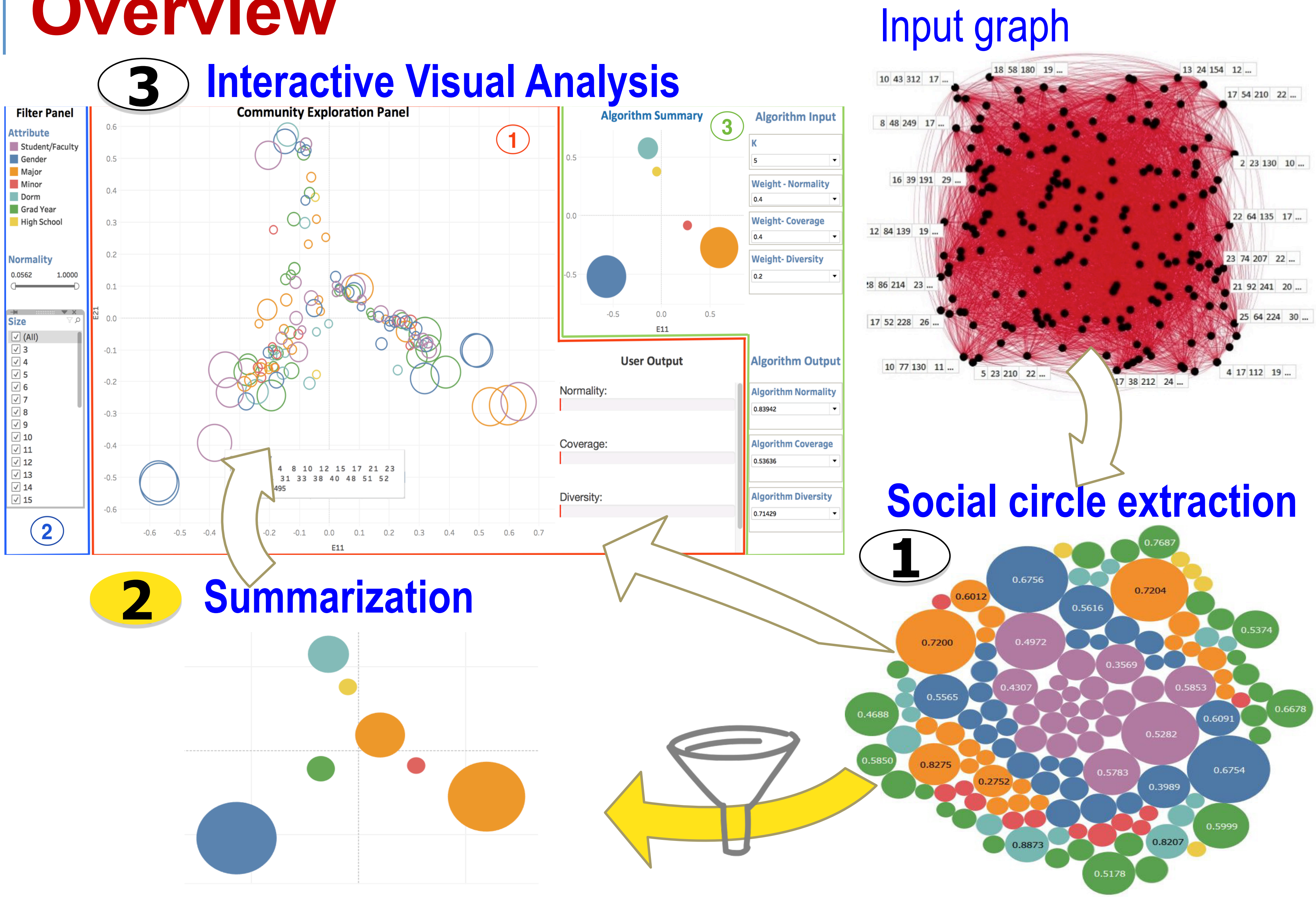
Algorithm 2 CONSTRUCTION *{build initial subgraph}*

Input: seed node s, $G = (\mathcal{V}, \mathcal{E}, \mathcal{A})$, $\mathbf{x}_{u \in \mathcal{V}}$, α

Output: initial subgraph S

$S = s$
while true **do**
  B := boundary nodes of S
  for each $b \in B$ **do**
    $\Delta N_b := N(S \cup b) - N(S)$
  end for
  if $\Delta N_b \leq 0, \forall b \in B$ **then return** S
  $max\Delta$:= maximum ΔN_b
  $min\Delta$:= minimum positive ΔN_b
  B_{cand} := boundary nodes for which:
    $\Delta N_b \geq min\Delta + \alpha * (max\Delta - min\Delta)$
  pick $v \in B_{cand}$ at random
  $S := S \cup v$
end while

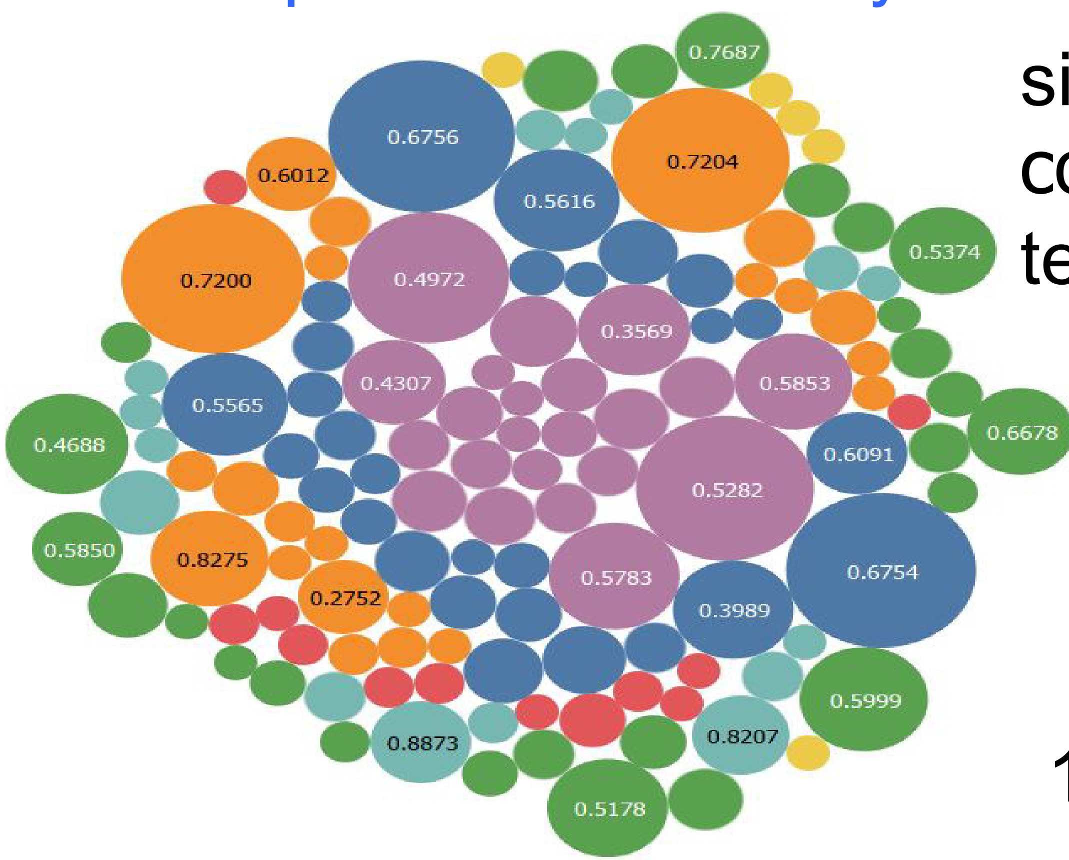
Overview



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)
Community Exploration Panel
Algorithm Summary
Algorithm Input
K
Weight - Normality
Weight- Coverage
Weight- Diversity
User Output
Normality:
Coverage:
Diversity:
Algorithm Output
Algorithm Normality
0.83942
Algorithm Coverage
0.53636
Algorithm Diversity
0.71429
E11
E21
Social circle extraction
1
2
Summarization

Summarization

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



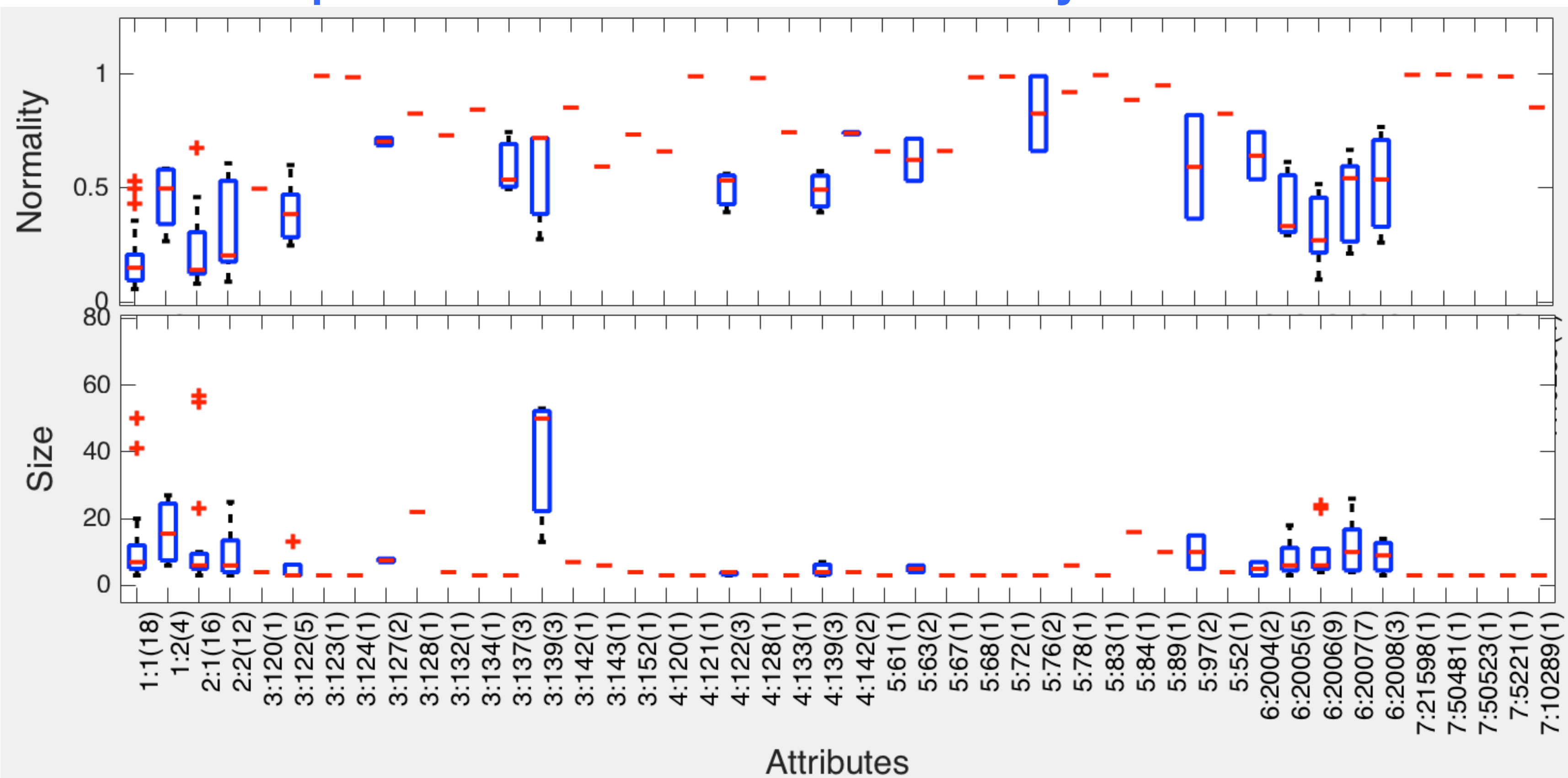
size $\propto$ #nodes
color: 'focus'
text: normality

125 circles!

  - does not reflect overlap between circles!

Summarization

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



  - does not reflect overlap between circles!

Summarization

- 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

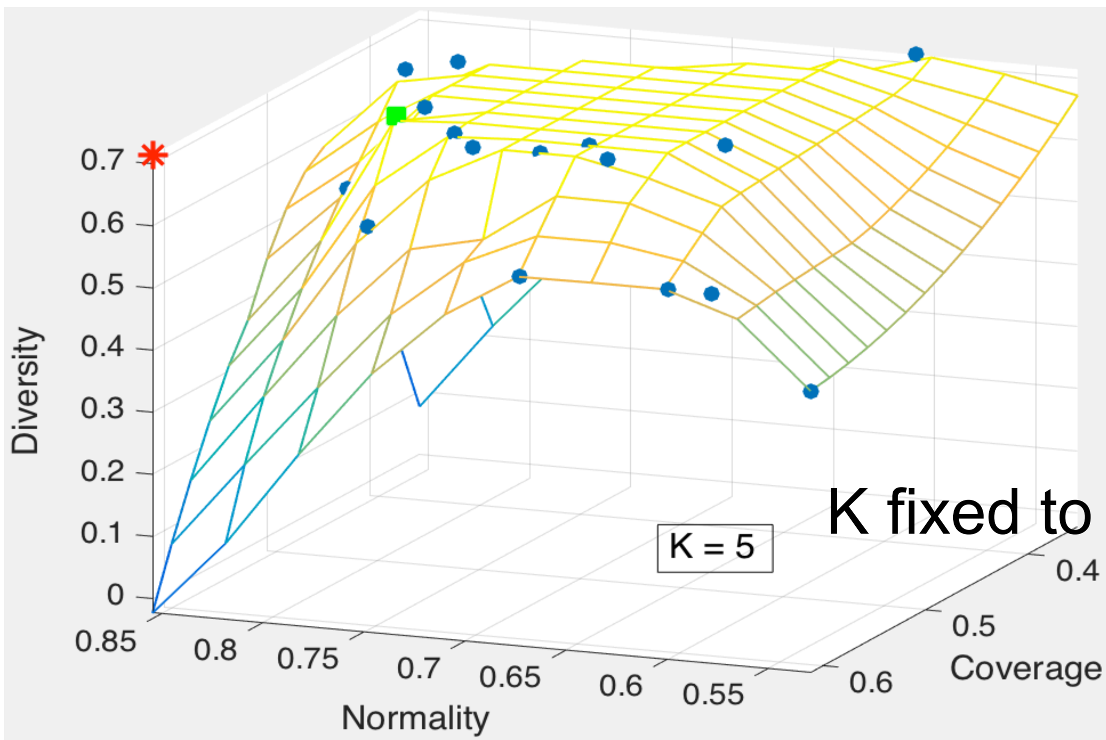
$$\max_{\substack{S \subseteq \mathcal{C} \\ |S|=K}} \quad f(S) = \alpha \underbrace{\frac{\sum_{C \in S} N(C)}{K}}_{\text{avg. normality}} + \beta \underbrace{\frac{|\bigcup_{C \in S} C|}{n}}_{\text{coverage}} + (1 - \alpha - \beta) \underbrace{\frac{|\bigcup_{C \in S} \mathcal{A}(C)|}{d}}_{\text{diversity}}$$

- 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

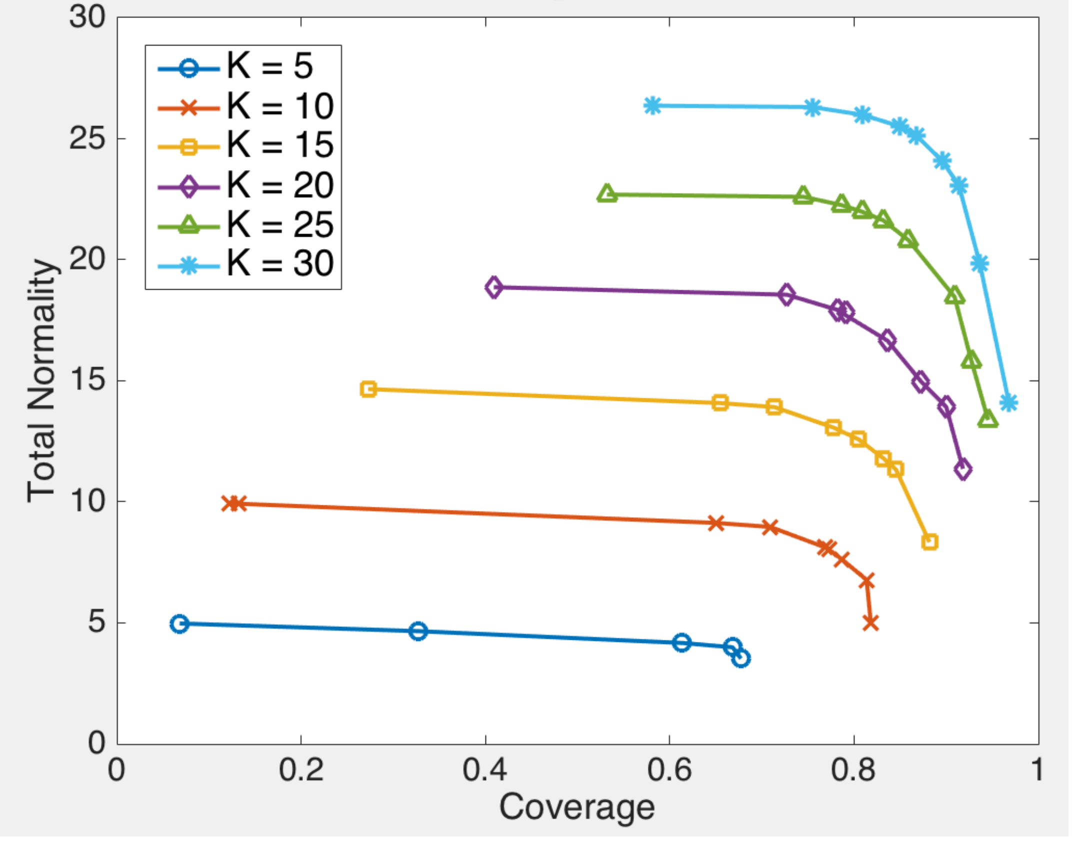
- surface formed by various parameter combinations $(\alpha, \beta, 1-\alpha-\beta)$ (blue dots)
  - (green) square around the "knee": a good trade-off between quality, coverage, and diversity



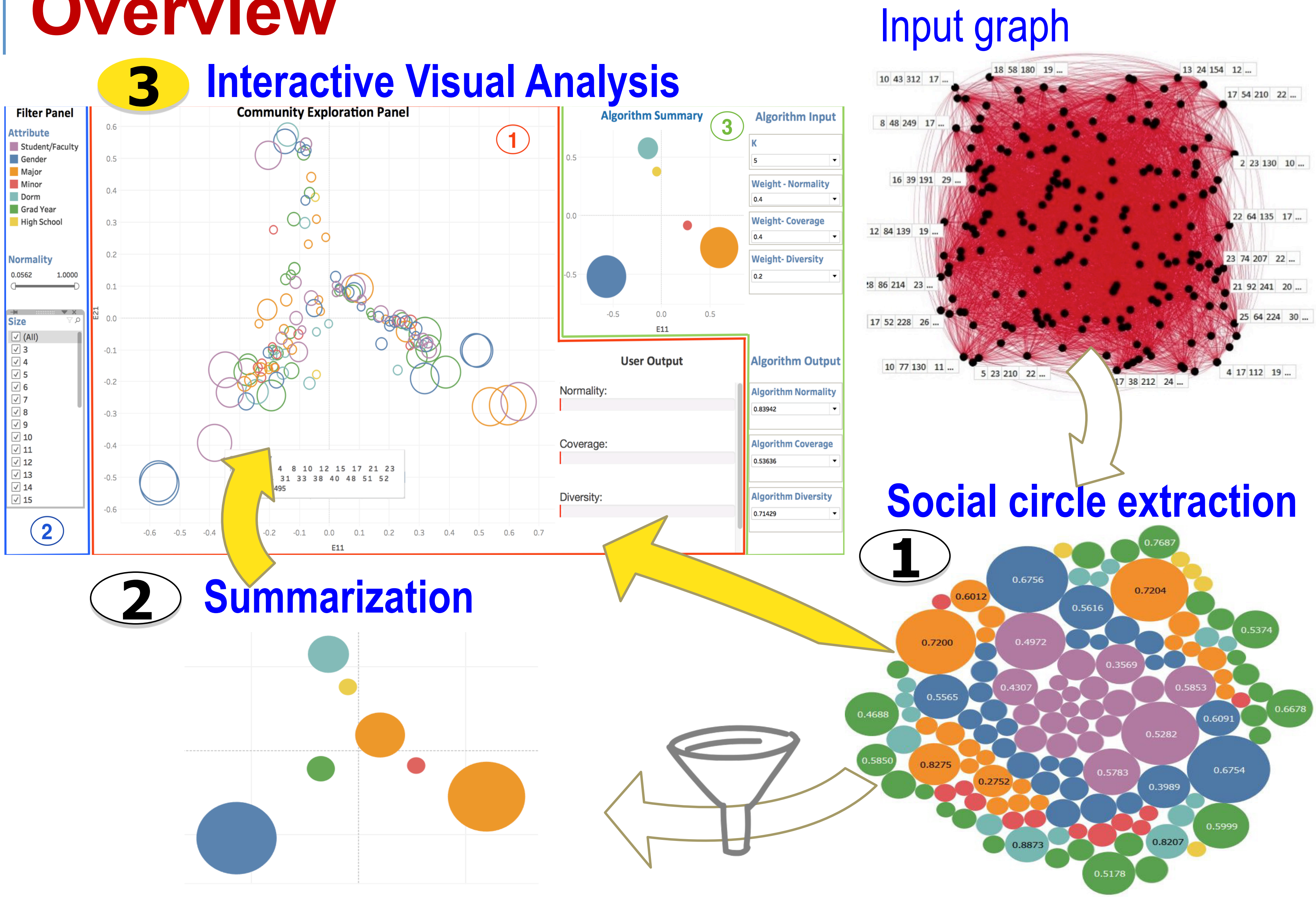
K fixed to 5

Summarization

- curves formed by various parameter combinations $\alpha + \beta = 1$ (diversity weight set to 0) for various K

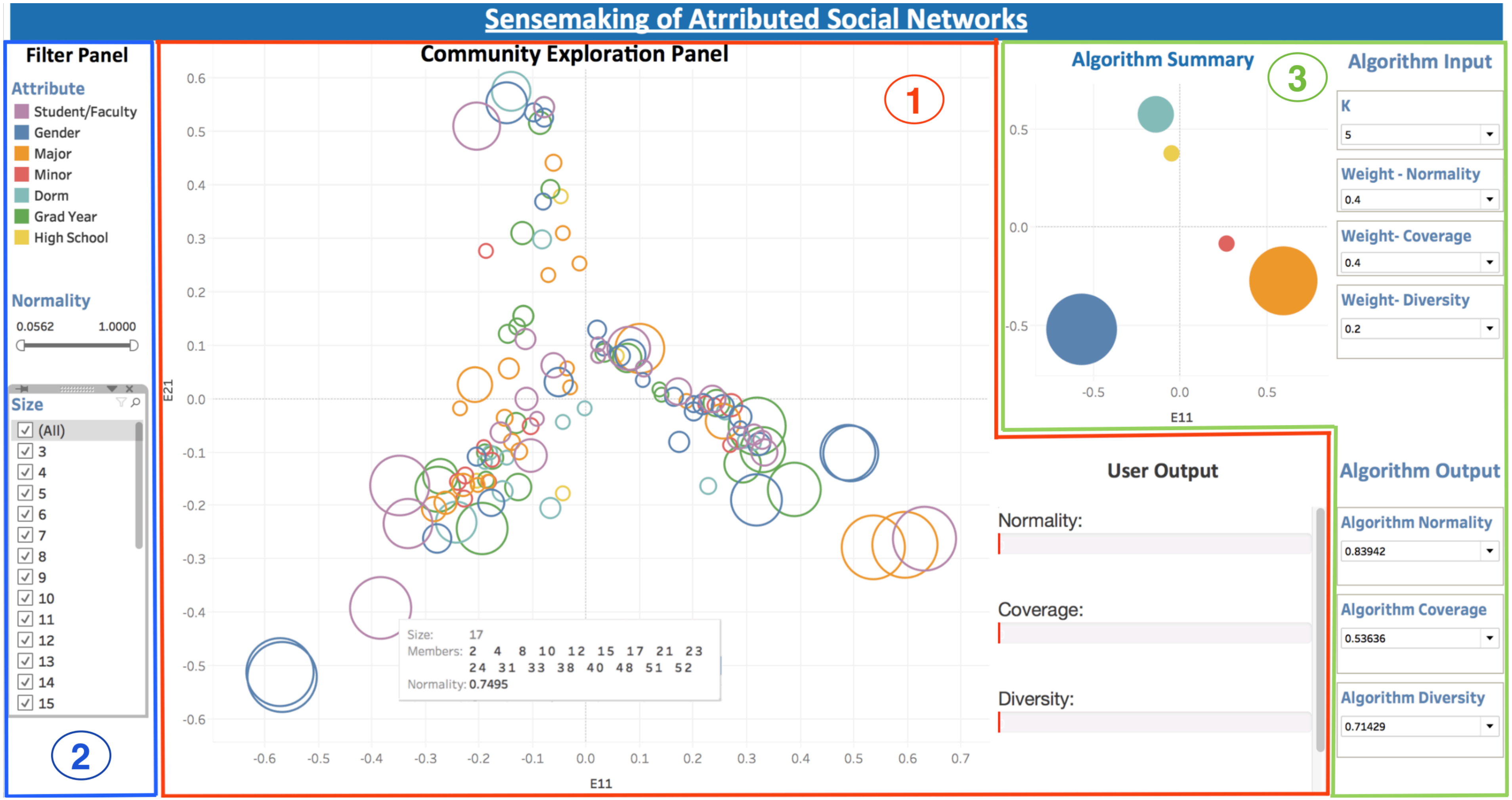


Overview

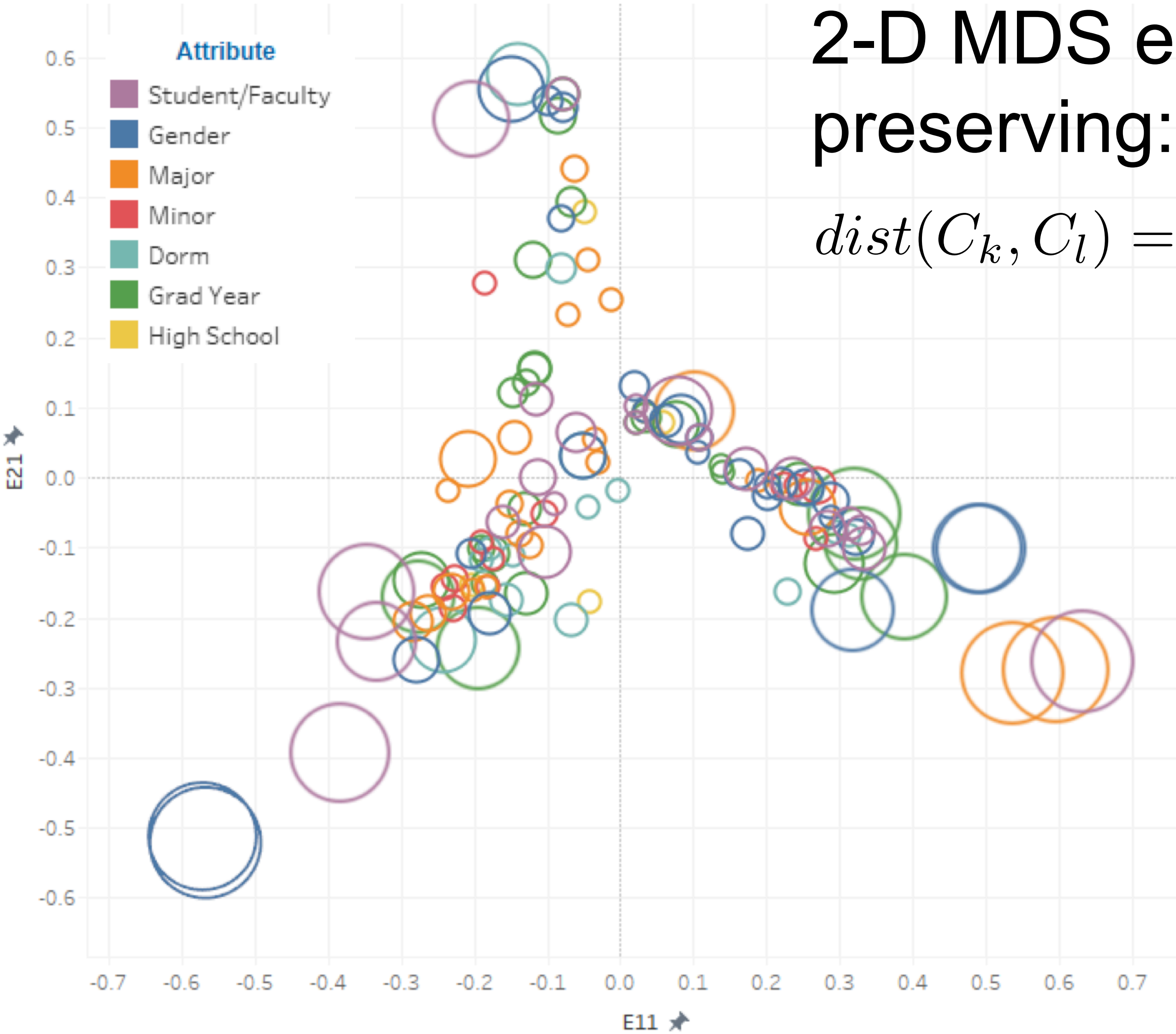


Input graph
3 Interactive Visual Analysis
Filter Panel
Attribute
Student/Faculty
Gender
Major
Minor
Dorm
Grad Year
High School
Normality
Size
Community Exploration Panel
Algorithm Summary
Algorithm Input
Weight - Normality
Weight- Coverage
Weight- Diversity
User Output
Normality:
Coverage:
Diversity:
Algorithm Output
Algorithm Normality
Algorithm Coverage
Algorithm Diversity
Social circle extraction
1
2 Summarization

Interactive Visual Exploration & Summarization



Circle exploration

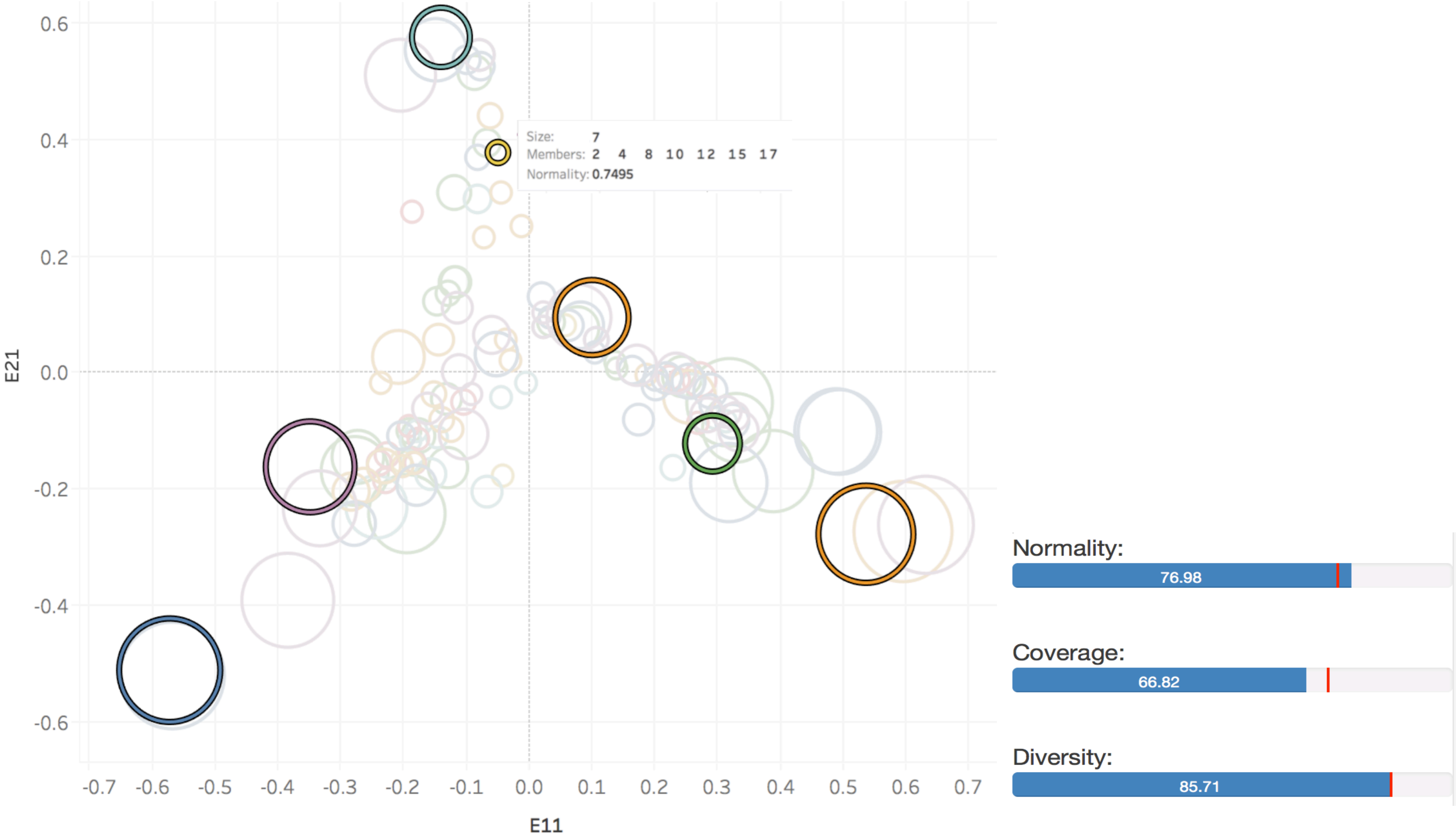


2-D MDS embedding preserving:

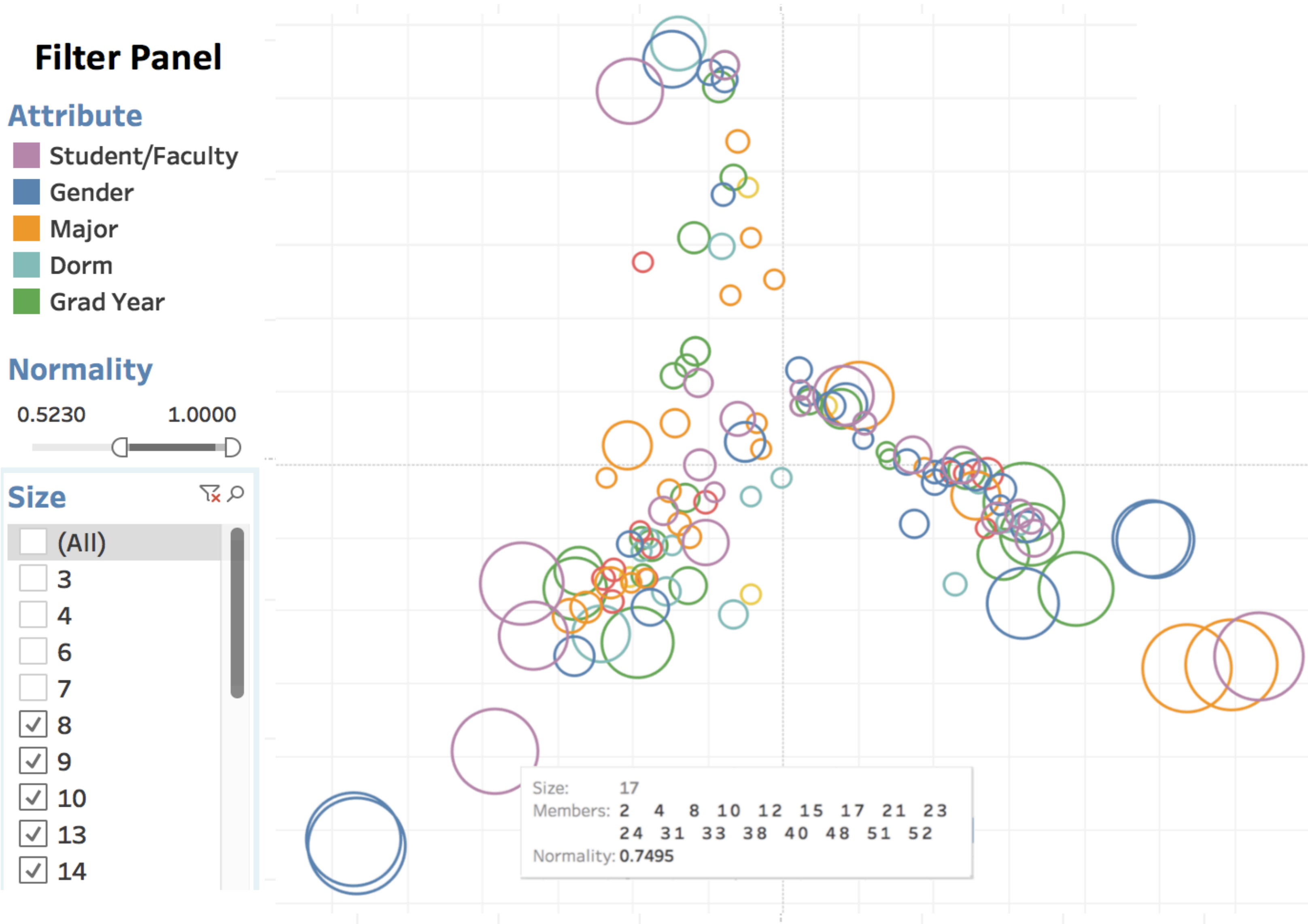
$$dist(C_k, C_l) = 1 - \frac{|C_k \cap C_l|}{\min(|C_k|, |C_l|)}$$

size $\propto$ #nodes
color: focus

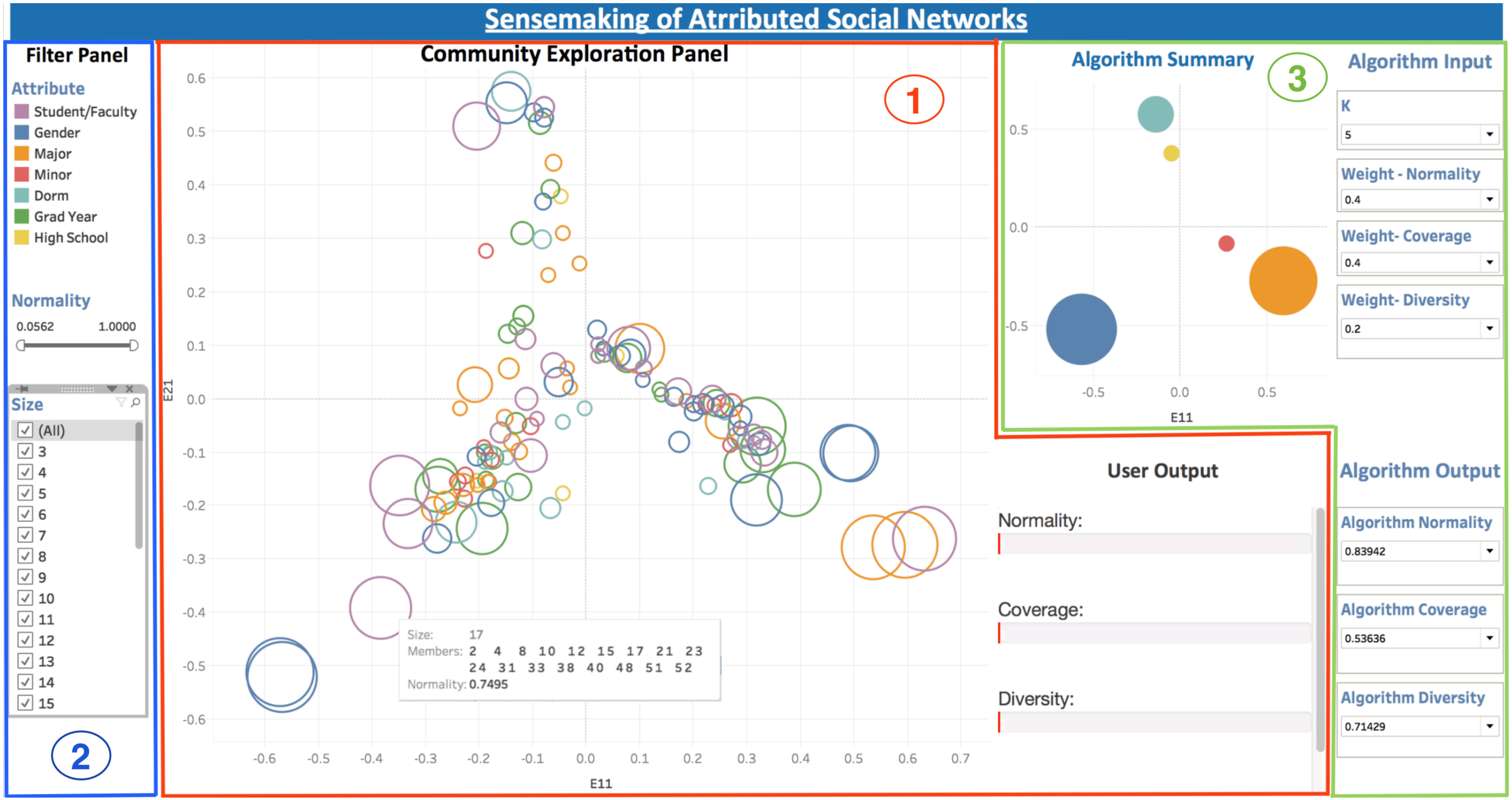
Circle exploration



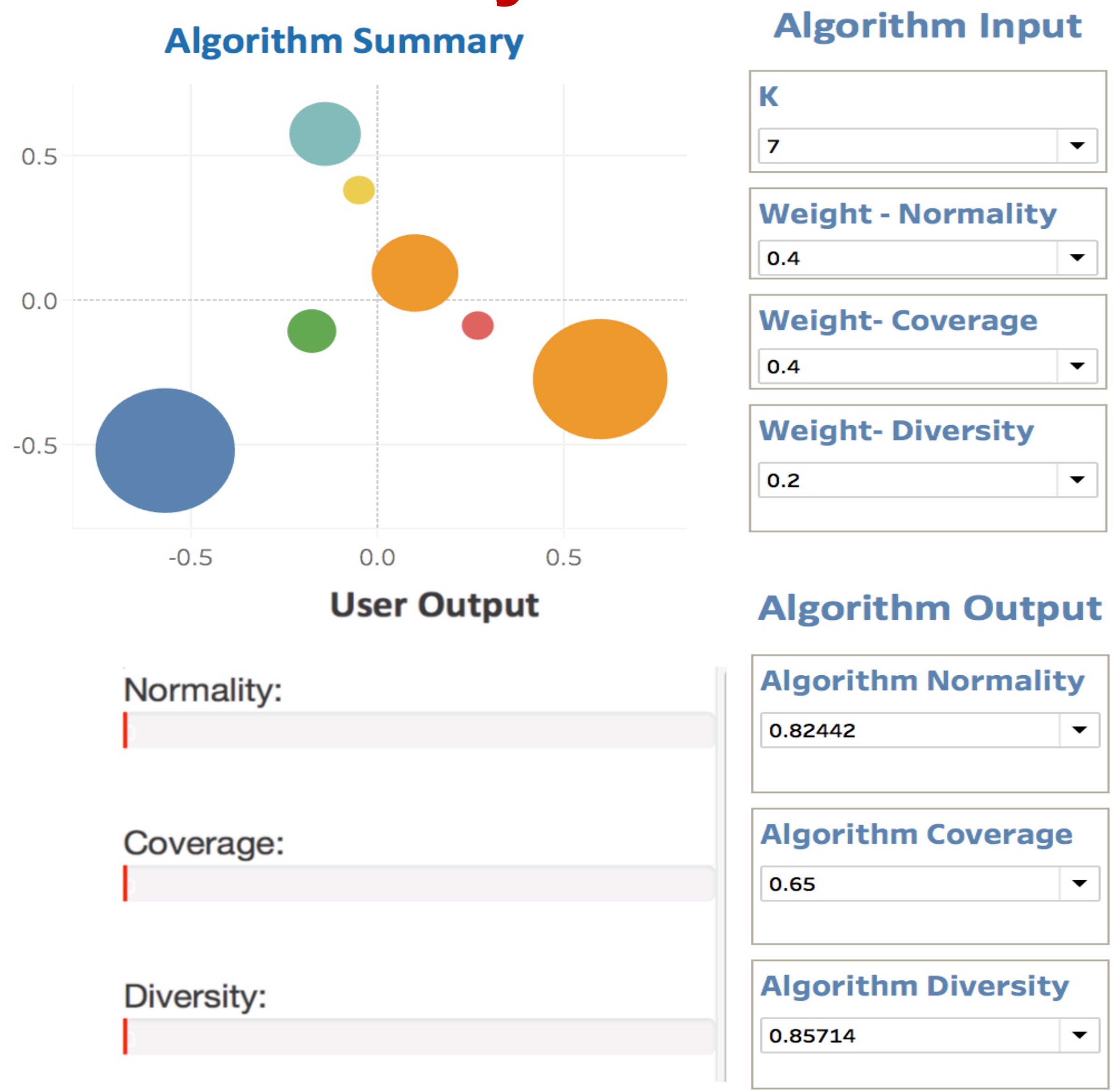
Filtering


Filter Panel
Attribute
Student/Faculty
Gender
Major
Dorm
Grad Year
Normality
0.5230
1.0000
Size
(All)
3
4
6
7
8
9
10
13
14
Size: 17
Members: 2 4 8 10 12 15 17 21 23
24 31 33 38 40 48 51 52
Normality: 0.7495

Interactive Visual Exploration & Summarization



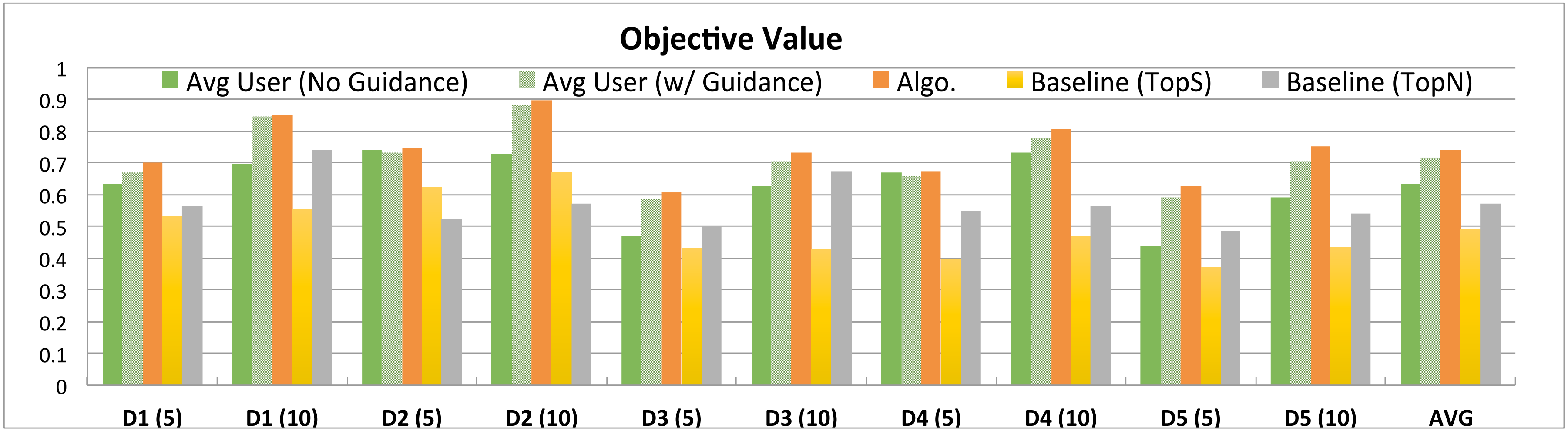
Algorithmic Summary



< Demo >

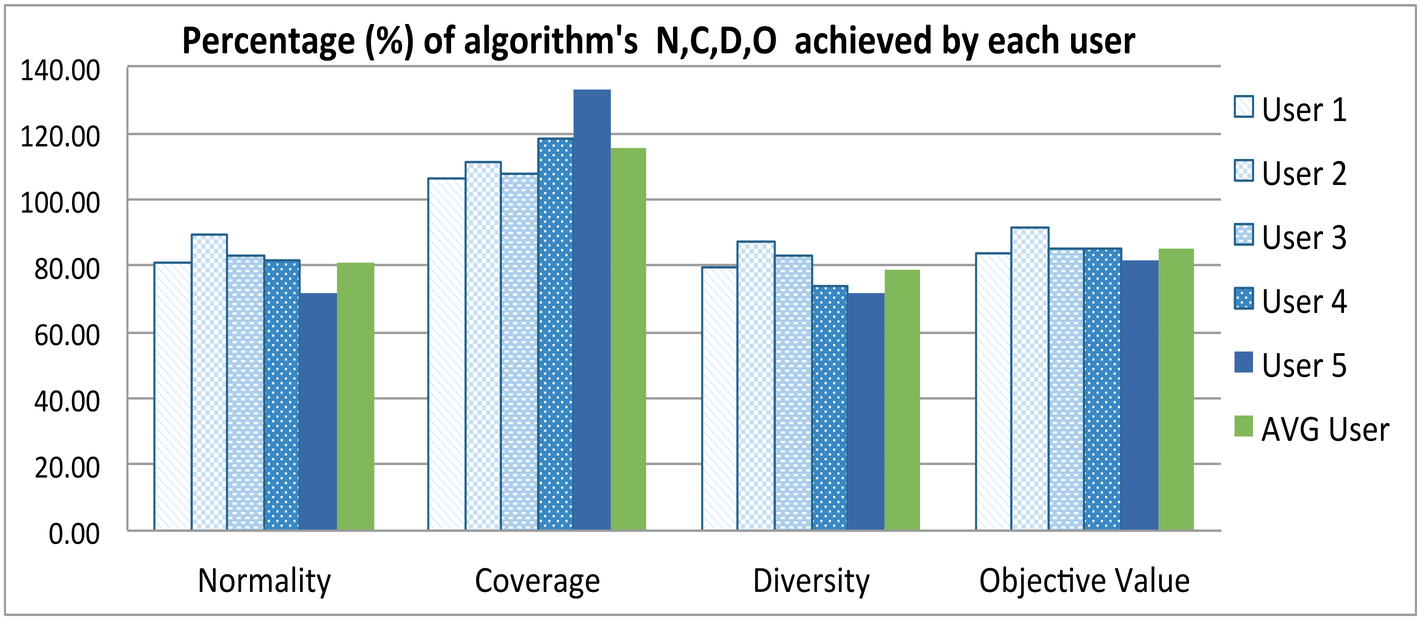
Evaluation

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



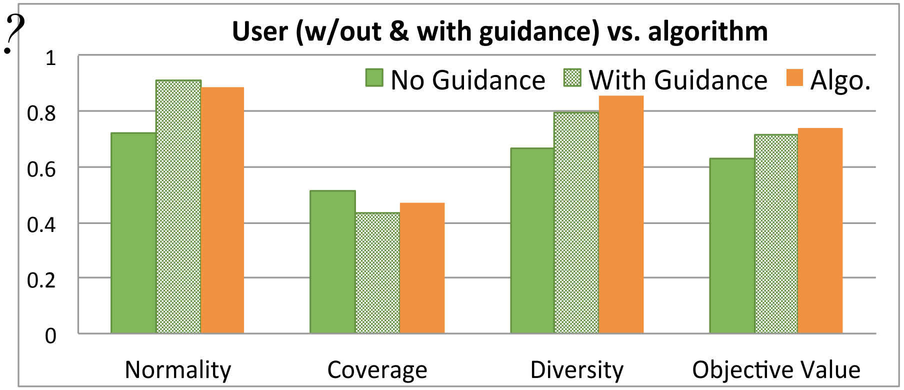
Evaluation

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?*



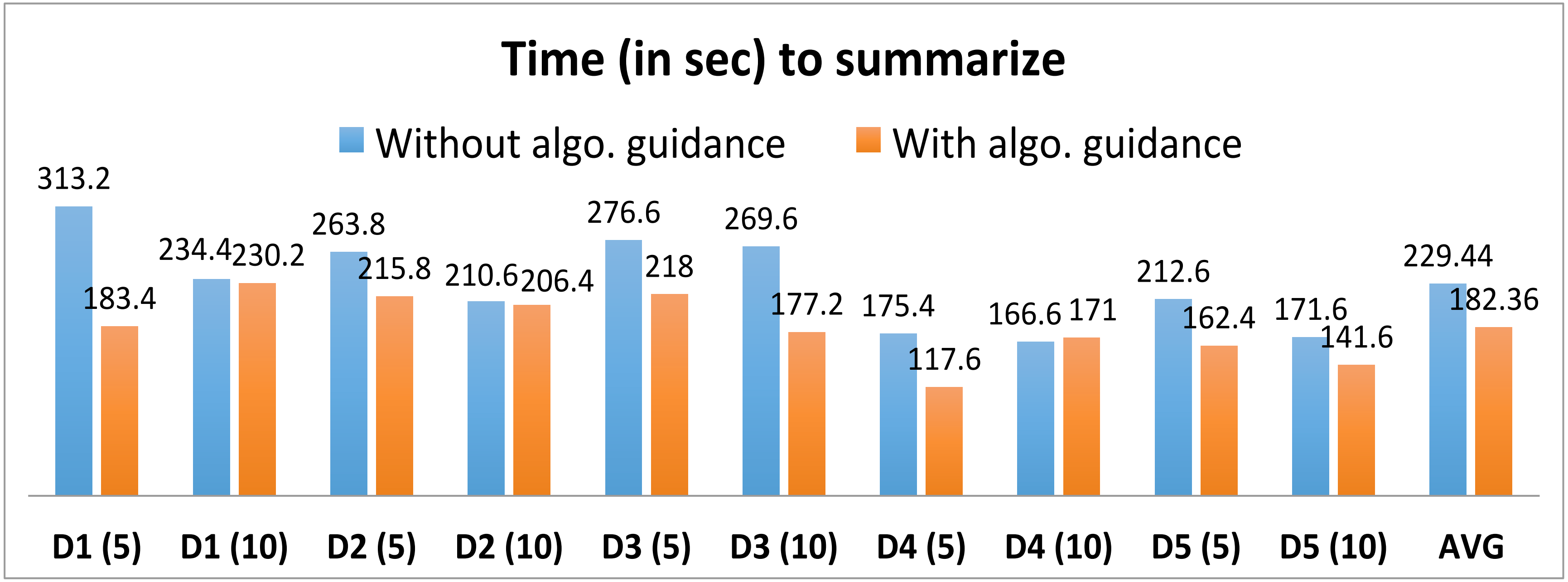
$$100\ O_{user}^{(after)} / O_{user}^{(before)}$$

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

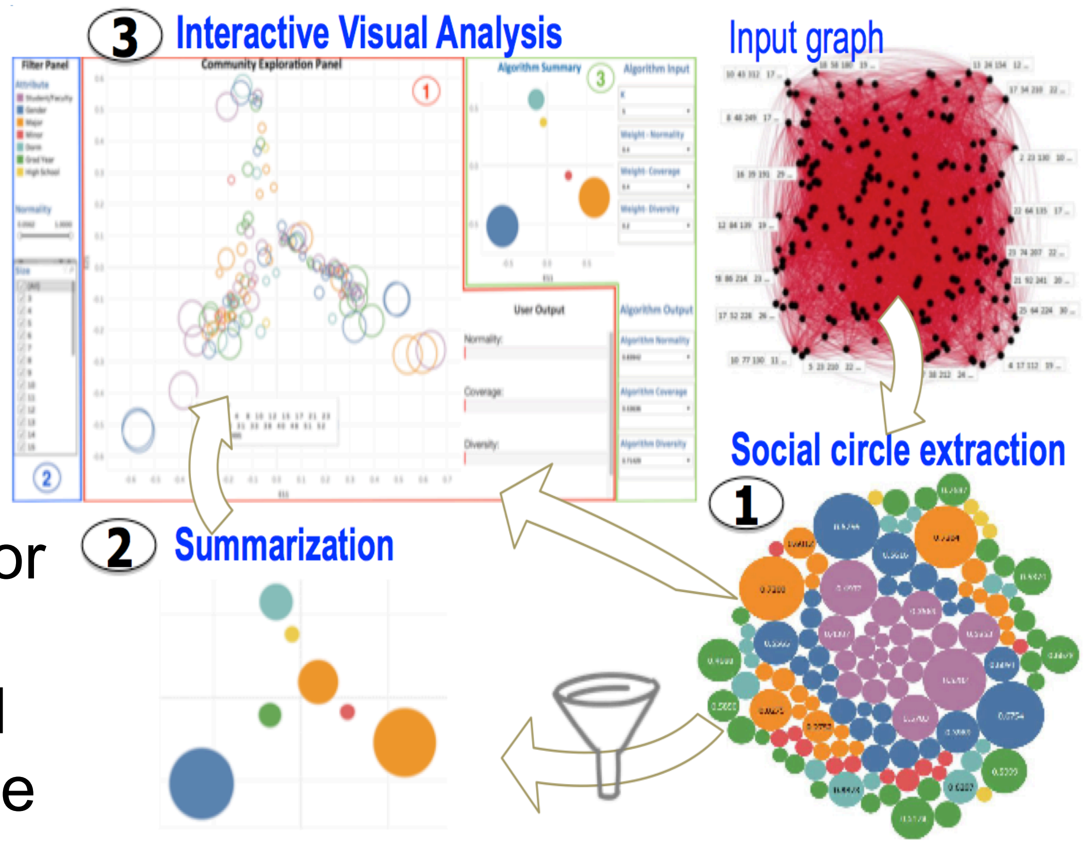
Q4) Efficiency. *How long does it take per user on average to construct (i) a summary without guidance, and (ii) alternative summary with guidance?*



Parting remarks

- An end-to-end system for sensemaking of node-attributed networks
  - main approach: “description-by-parts”

1. Normality measure & circle extraction
2. Summarization wrt
   - quality,
   - coverage, and
   - diversity
3. Interactive interface for
   - exploration,
   - summarization, and
   - algorithmic guidance



Thanks & Questions

lakoglu@cs.cmu.edu
www.cs.cmu.edu/~lakoglu