Focused Clustering and Outlier Detection in Large Attributed Graphs

ACM SIG-KDD
August 26, 2014

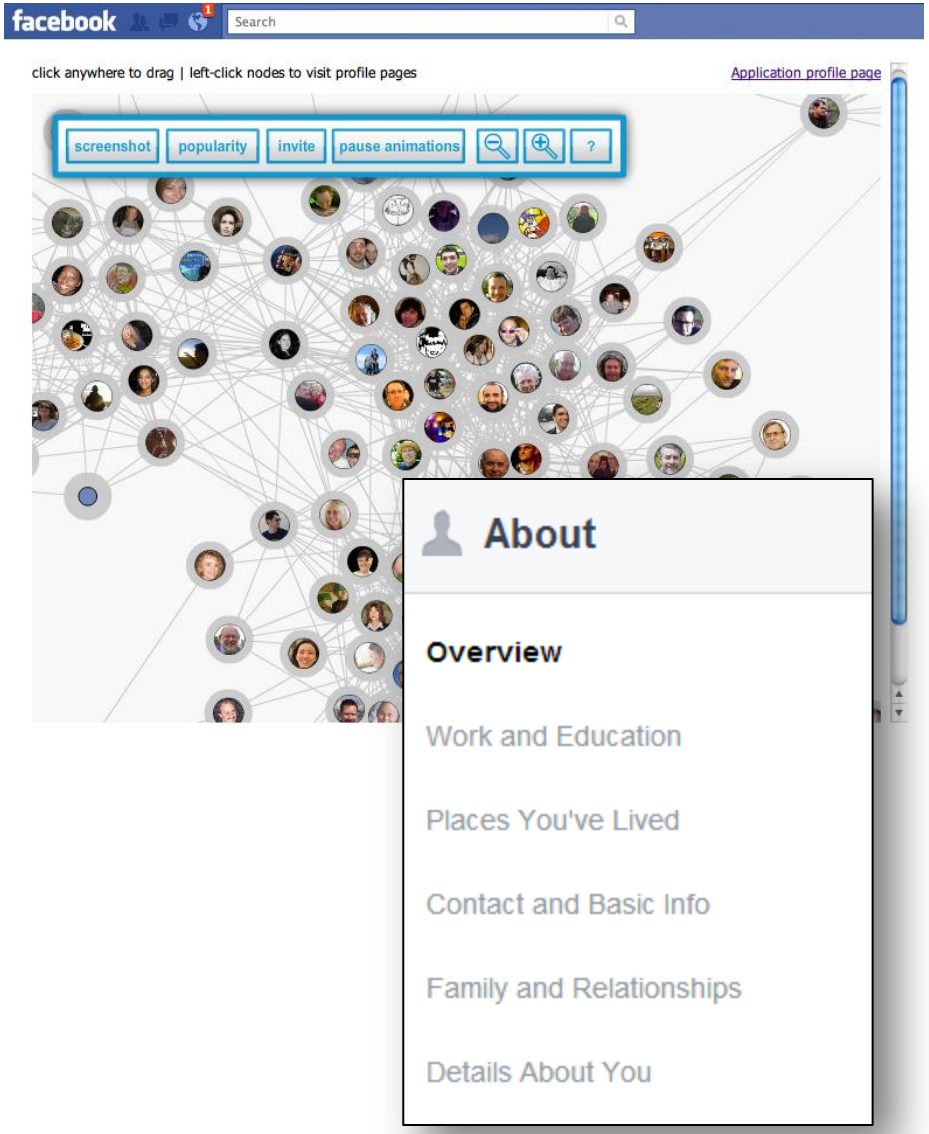

Bryan Perozzi, Leman Akoglu
Stony Brook University


Patricia Iglesias Sánchez*, Emmanuel Müller*†
*Karlsruhe Institute of Technology
†University of Antwerp

Stony Brook University
Computer Science
KIT Karlsruher Institut für Technologie
Universiteit Antwerpen

Attributed Graphs

- Attributed graph: each node has 1+ properties
- Examples:
  - Age
  - School
  - Relationship Status



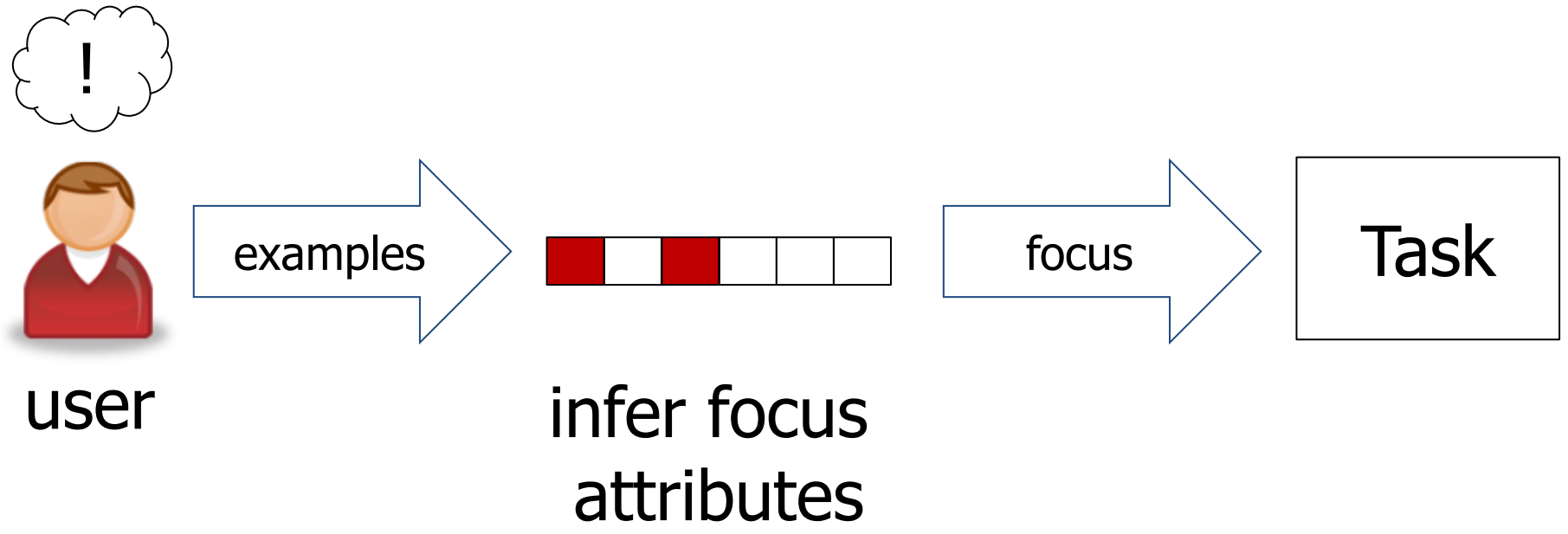
Focused Mining of Attributed Graphs

- Numerous attributes (ex: Facebook profiles)
- Many irrelevant for most queries
  - Ex: When trying to sell mortgages
    - **Useful**: Income, Credit Score, Employer ← **Focus**
    - **Not Useful**: Hair Color, # Apps Installed
  - Ex: When trying to sell make up
    - **Useful**: Hair Color, Skin Tone, Gender ← **Focus**
    - **Not Useful**: Shoe Size

Users have a Focus → Algorithms need a Focus too!

Adding Focus to Algorithms

- Users provide examples of the kind of similarity they are interested in.
- We infer the similarity function that matters to them.



Outline

- Introduction
- **New Problem:**
  Focused Clustering & Outliers
- Our Approach: FocusCO
- Evaluation
- Conclusion

Focused Clusters and Outliers: Problem

Given

1) a graph w/ node attributes,
2) exemplar nodes by the user

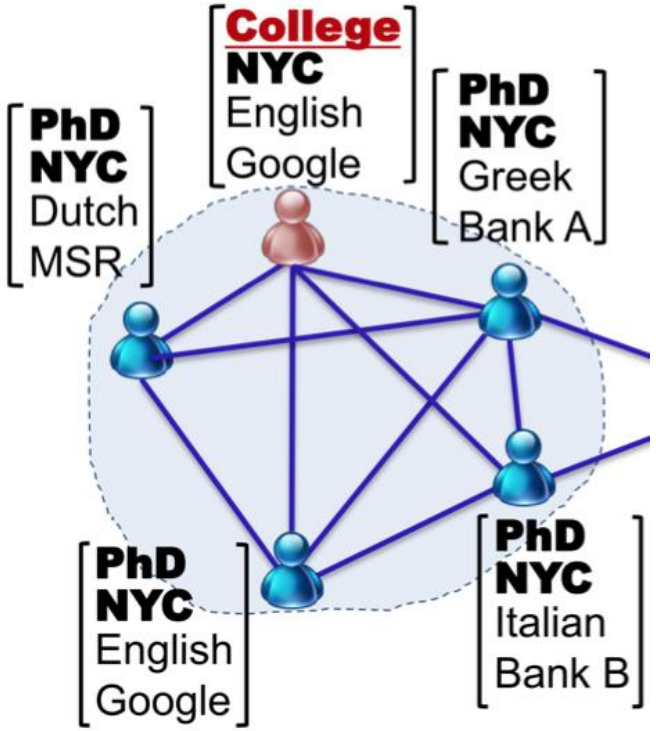
Infer attribute weights/relevance

Extract focused clusters:

1) dense in structure,
2) coherent in "heavy" attributes
   (called the "focus")

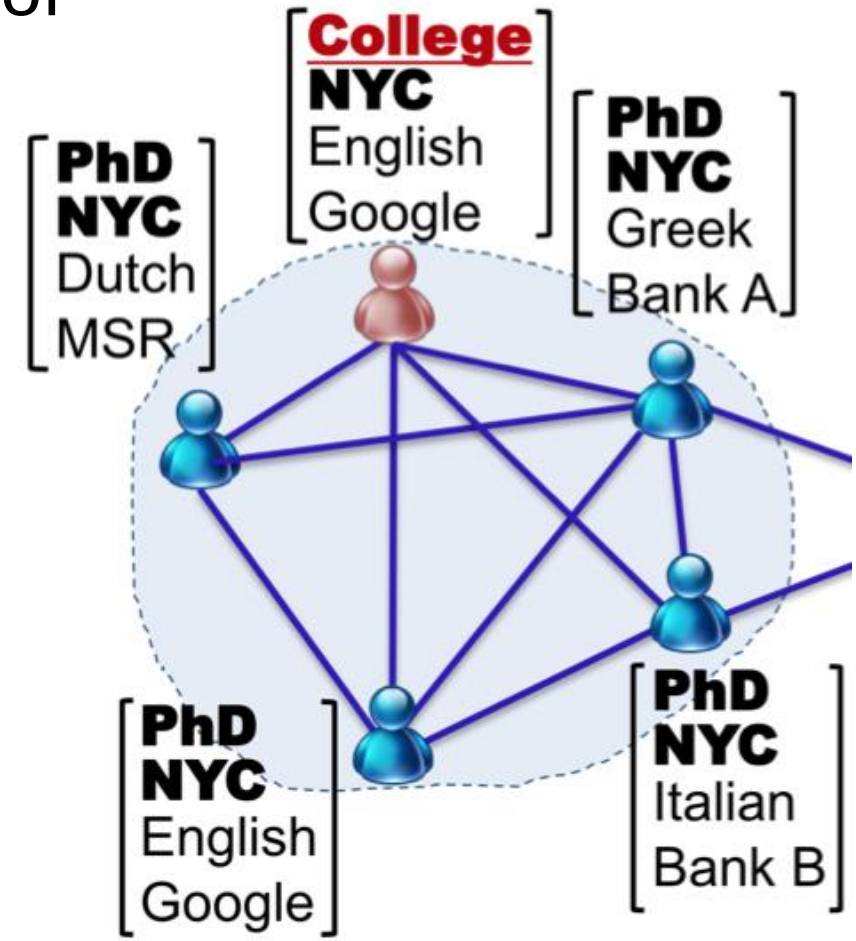
Detect focused outliers:

*) nodes deviating in focus attribute values



An Example

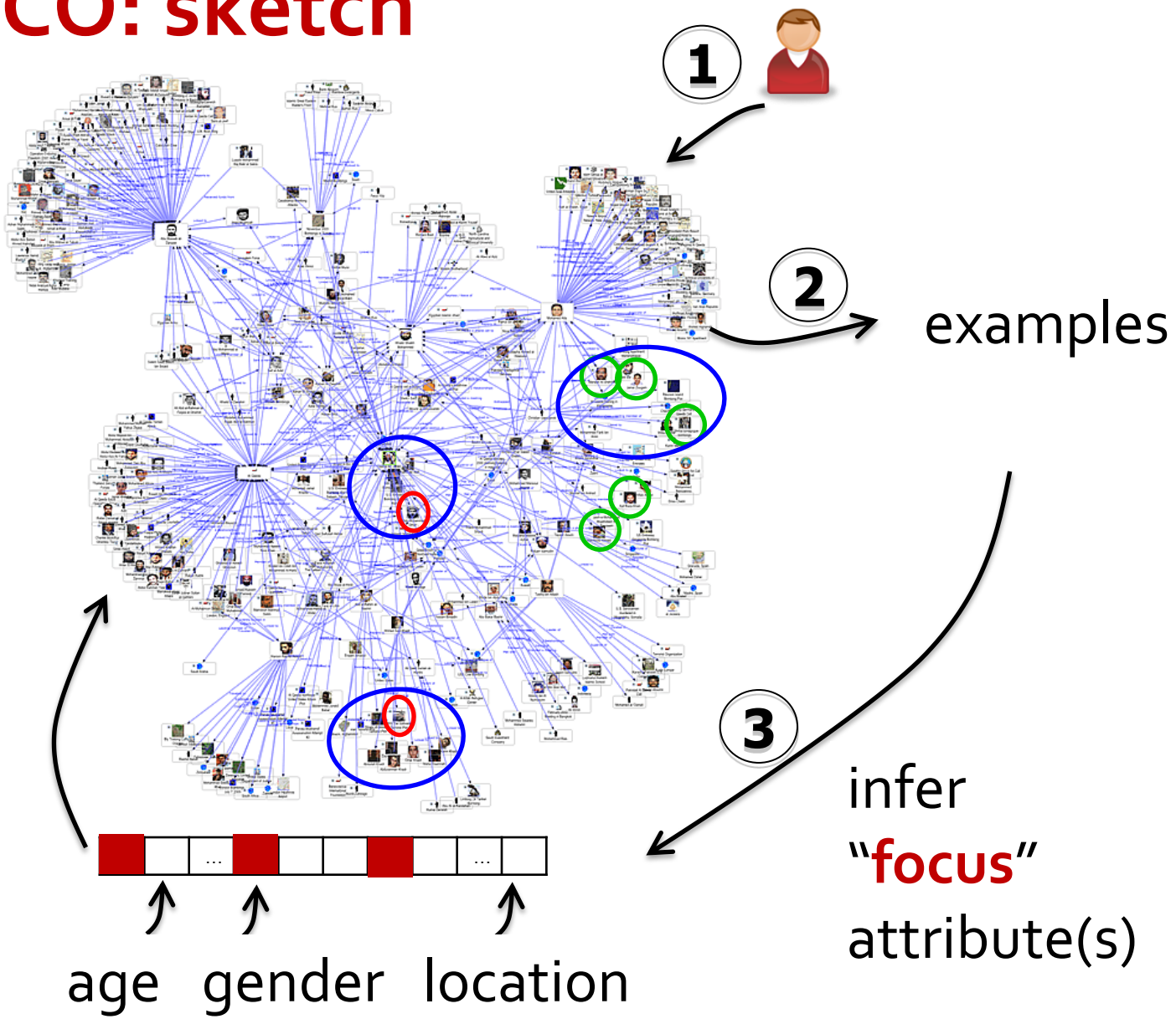
- Users provide examples of nodes they consider similar.
  - Ex: ‘Yann LeCun’ and ‘Foster Provost’
- We learn a focus
  - Education Level
  - Location
- We extract clusters
  - which agree with the focus
- We detect outliers
  - which don’t agree with focus



Related Work

	Graph Clustering	Attributed Graphs	Attribute Subspace	User Preference	Outlier Detection
METIS, Spectral	✓				
Parallel Nibble, BigClam	✓				
CoPaM, Gamer	✓	✓	✓		
CODA	✓	✓			✓
GOutRank, ConSub		✓	✓		✓
FocusCO	✓	✓	✓	✓	✓

FocusCO: sketch

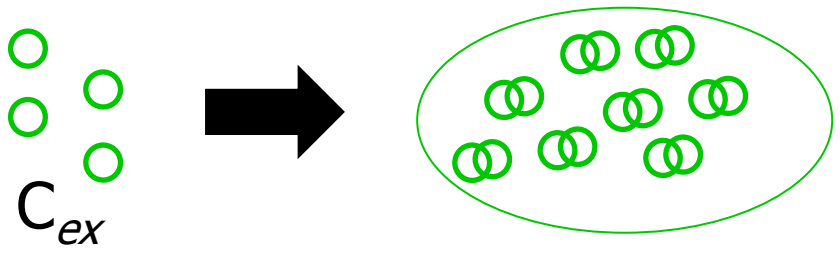


Focus attribute inference

Input: Set of similar nodes, C_{ex}

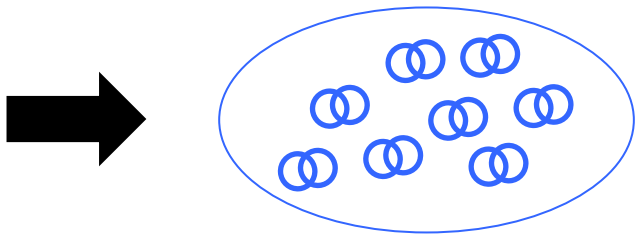
1. Construct a set of similar pairs, P_S

Pair user examples together



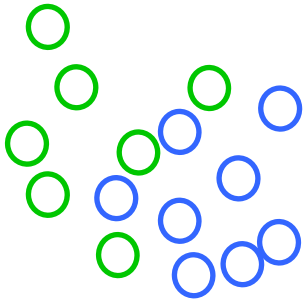
2. Construct a set of dissimilar pairs, P_D

Randomly sample pairs (u,v)

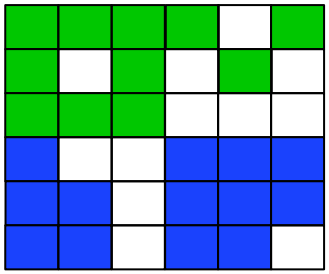


3. Learn a distance metric between P_S and P_D

Distance Metric Learning [Xing, et al 2002]



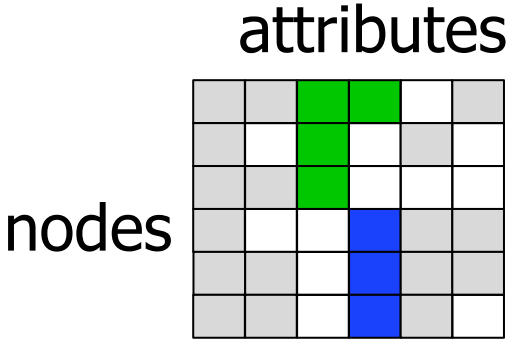
P_S and P_D intermixed

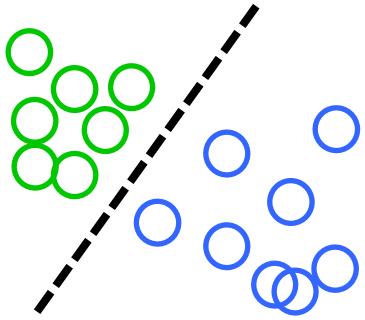


Feature Matrix

$$\min_{\boldsymbol{A}} \sum_{(i,j)\in P_S} (\boldsymbol{f_i} - \boldsymbol{f_j})^T \boldsymbol{A} (\boldsymbol{f_i} - \boldsymbol{f_j}) - \gamma \log \left(\sum_{(i,j)\in P_D} \sqrt{(\boldsymbol{f_i} - \boldsymbol{f_j})^T \boldsymbol{A} (\boldsymbol{f_i} - \boldsymbol{f_j})} \right)$$

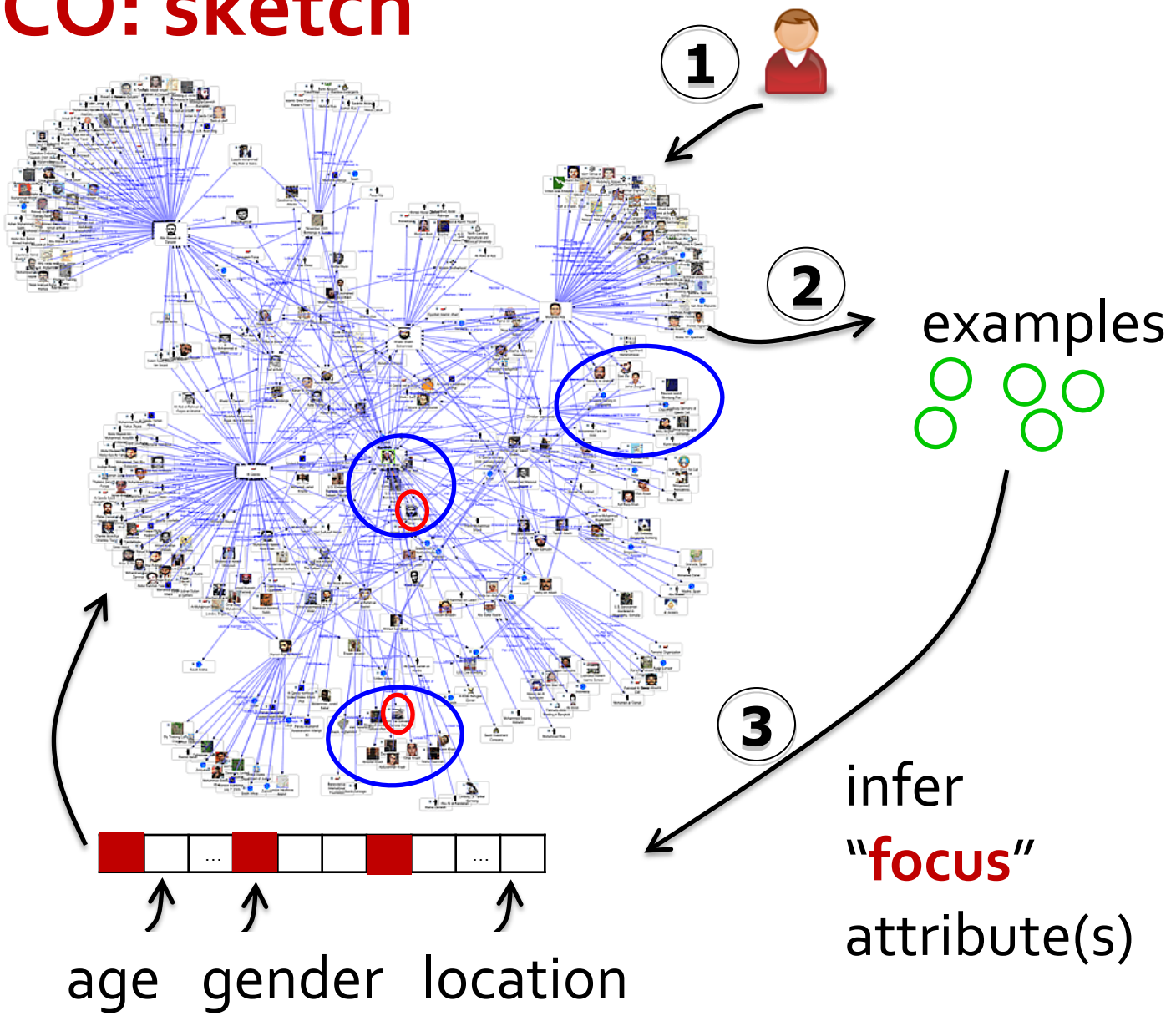
Focused Attribute Vector





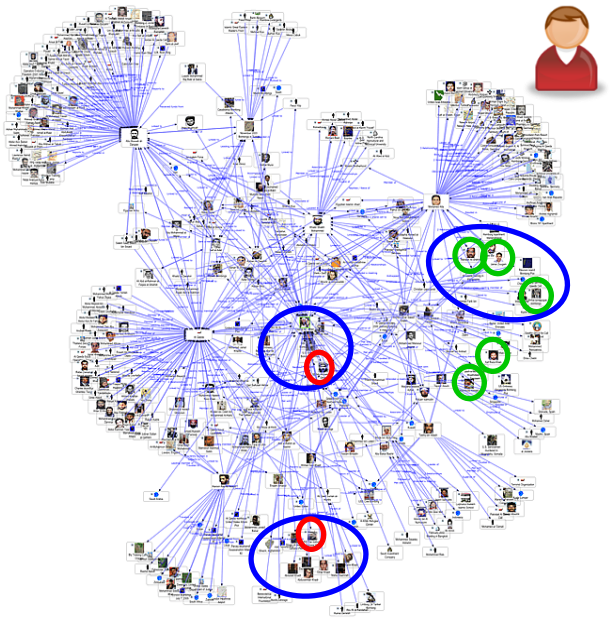
P_S closer together

FocusCO: sketch



FocusCO: Cluster Extraction

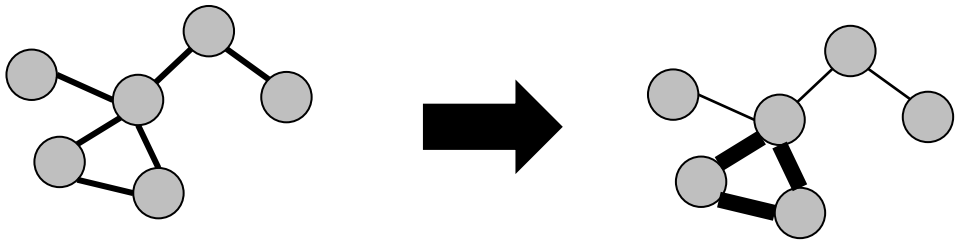
- Local clustering algorithm
  - Not cluster whole graph
- Expands a cluster around a starting set
- Two procedures:
  1. Finding good candidate sets to start at
  2. Growing clusters



Finding nodes to cluster around

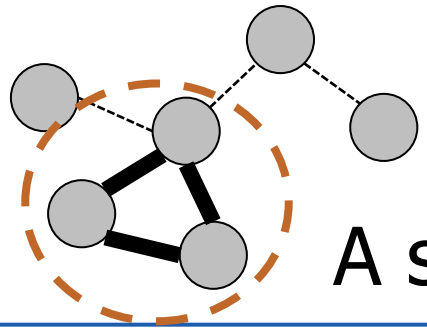
1.) We reweigh the graph using the focus

$$\textbf{for each } (i,j) \in E \textbf{ do}$$
$$w(i,j) = 1/(1+\sqrt{(\boldsymbol{f_i}-\boldsymbol{f_j})^T diag(\boldsymbol{\beta})(\boldsymbol{f_i}-\boldsymbol{f_j})})$$



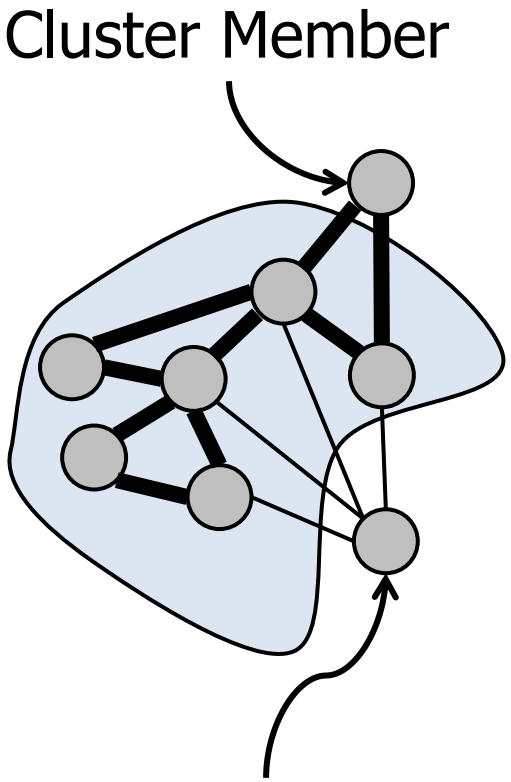
2.) We keep only highly weighted edges

3.) The connected components are our seeds



A seed set

Growing a Focused Cluster



1. Clustering objective: conductance $\phi^{(w)}$ weighted by focus

$$\phi^{(w)}(C, G) = \frac{W_{cut}(C)}{WVol(C)}$$

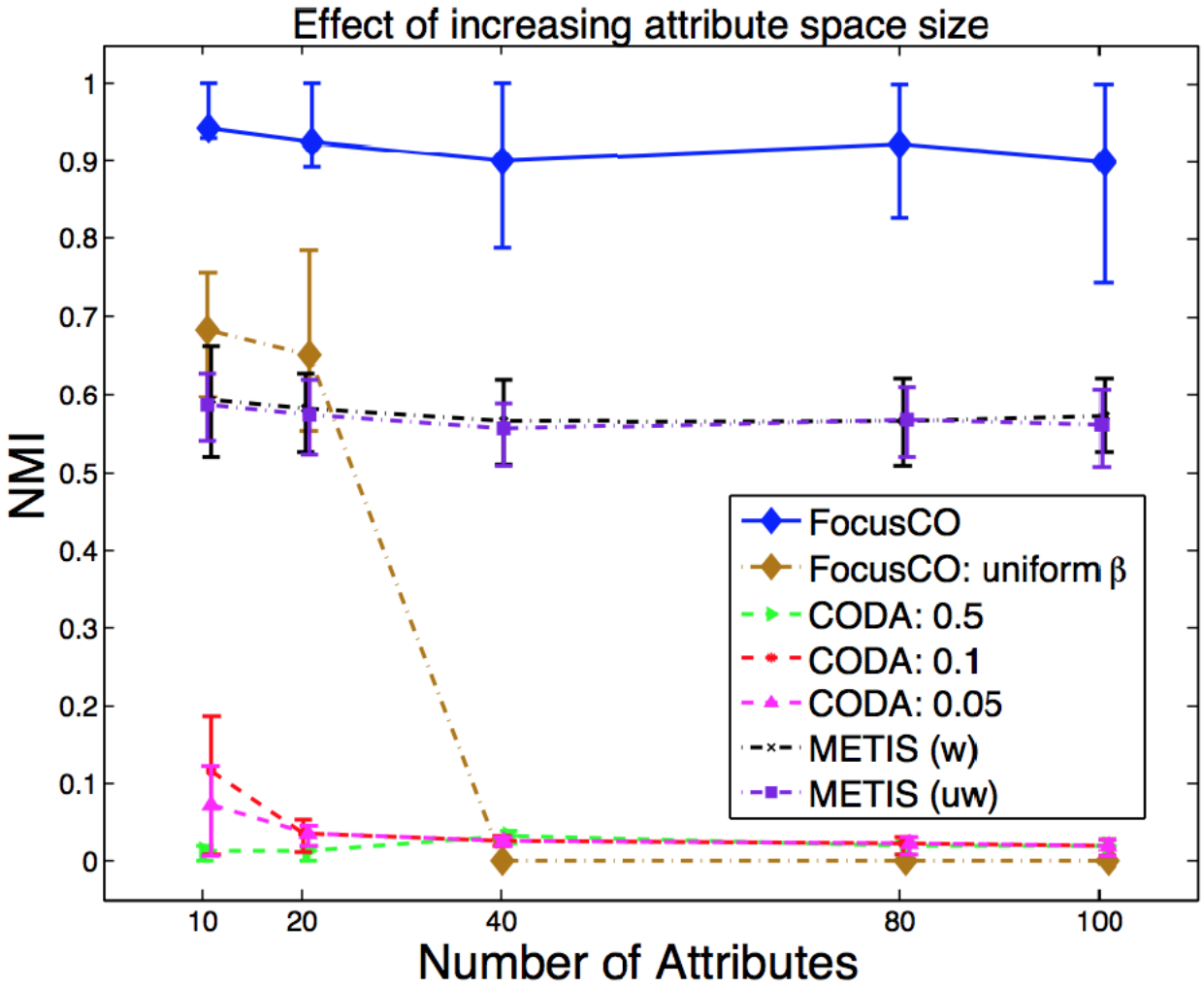
2. At each step in cluster expansion:
   - 2.1 - Examine boundary nodes
   - 2.2 - Add node with best $\Delta\phi^{(w)}$
   - 2.3 - Record best structural node

3. Focused Outliers:
   - left out best structural nodes

Experiment set up

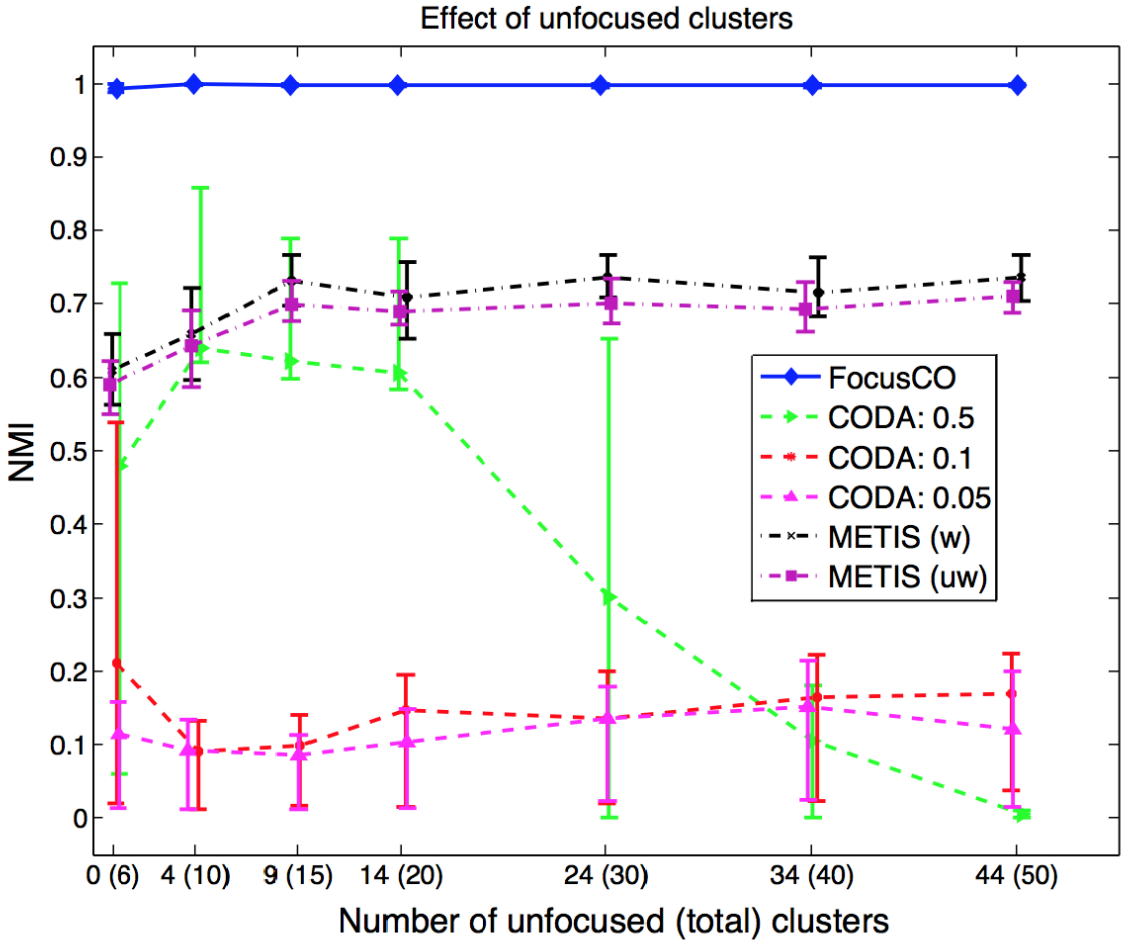
- Synthetic and Real World Graphs
- Performance measures:
  - Cluster quality: NMI
  - Outlier accuracy: precision, F1
- Compared to:
  - CODA [Gao+'10]
  - METIS (no outlier detection) [Karypis+'98]

Focused clustering performance



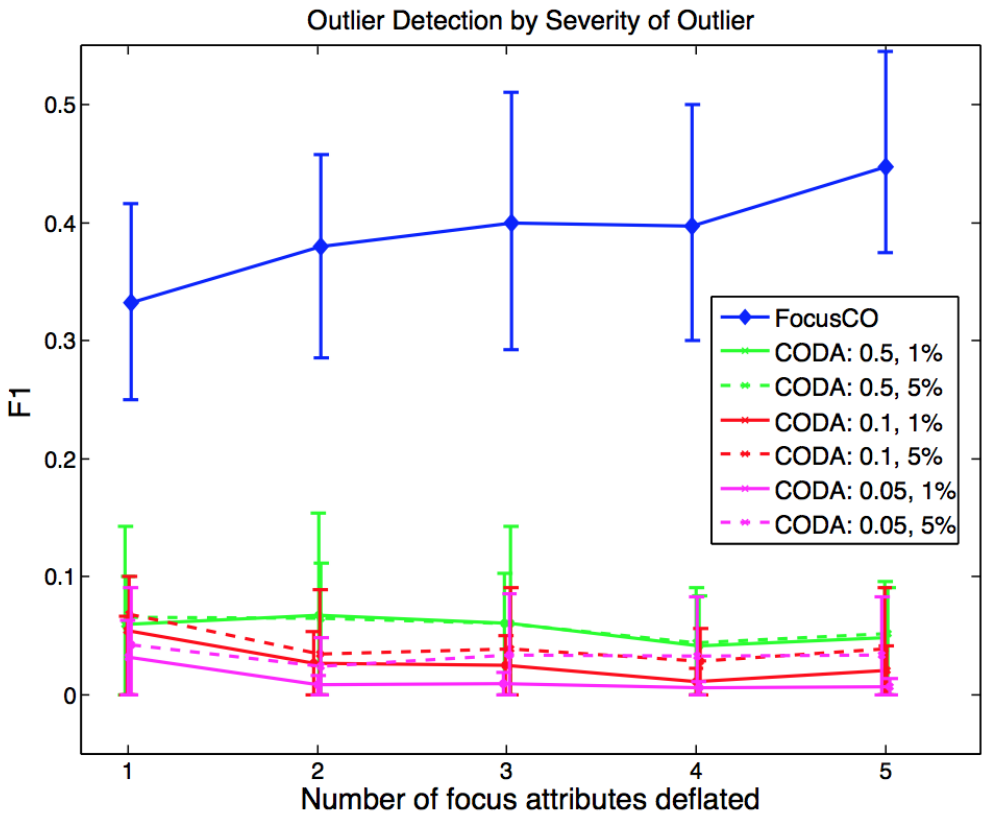
9 clusters (3 focus1 + 3 focus2 +3 unfocused). 5 focus attributes.

Focused clustering performance

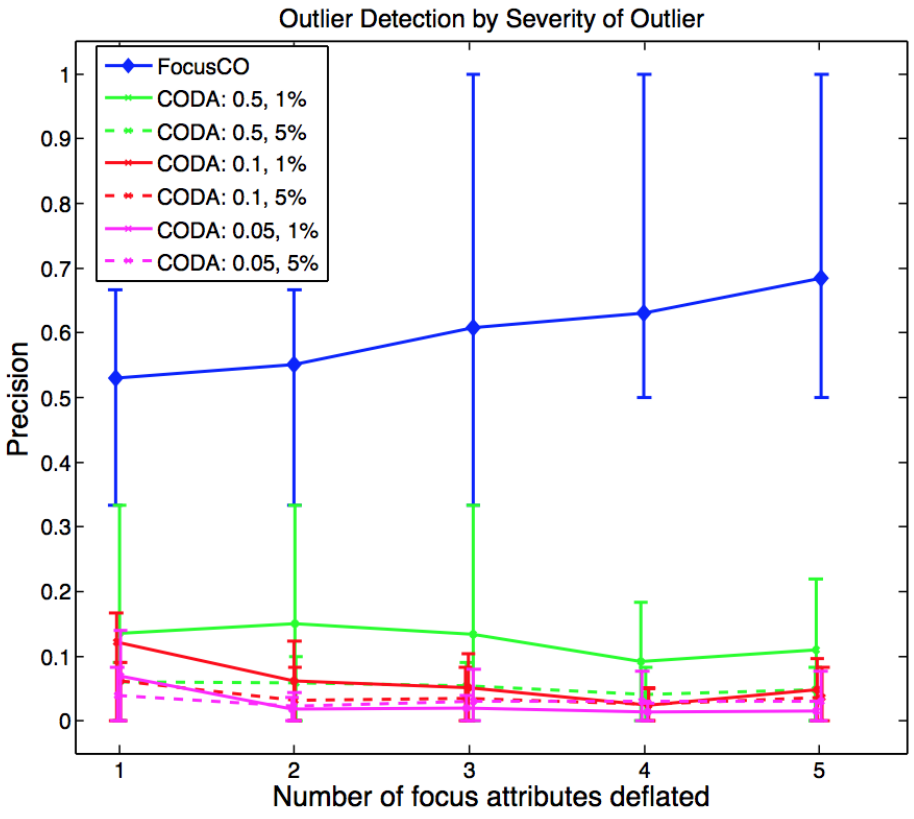


(c) NMI vs. number of unfocused clusters

Outlier detection performance



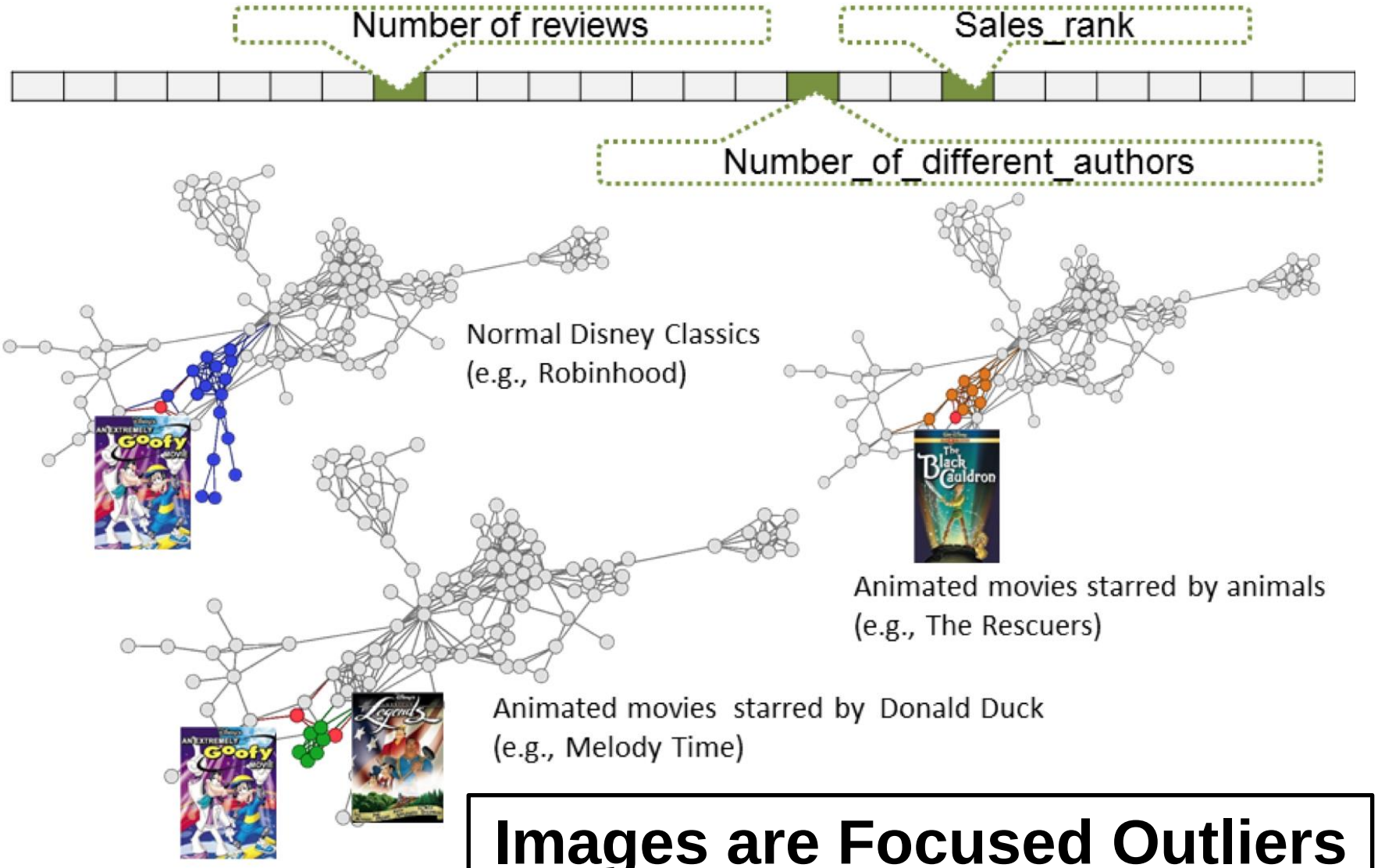
(a) F1-score



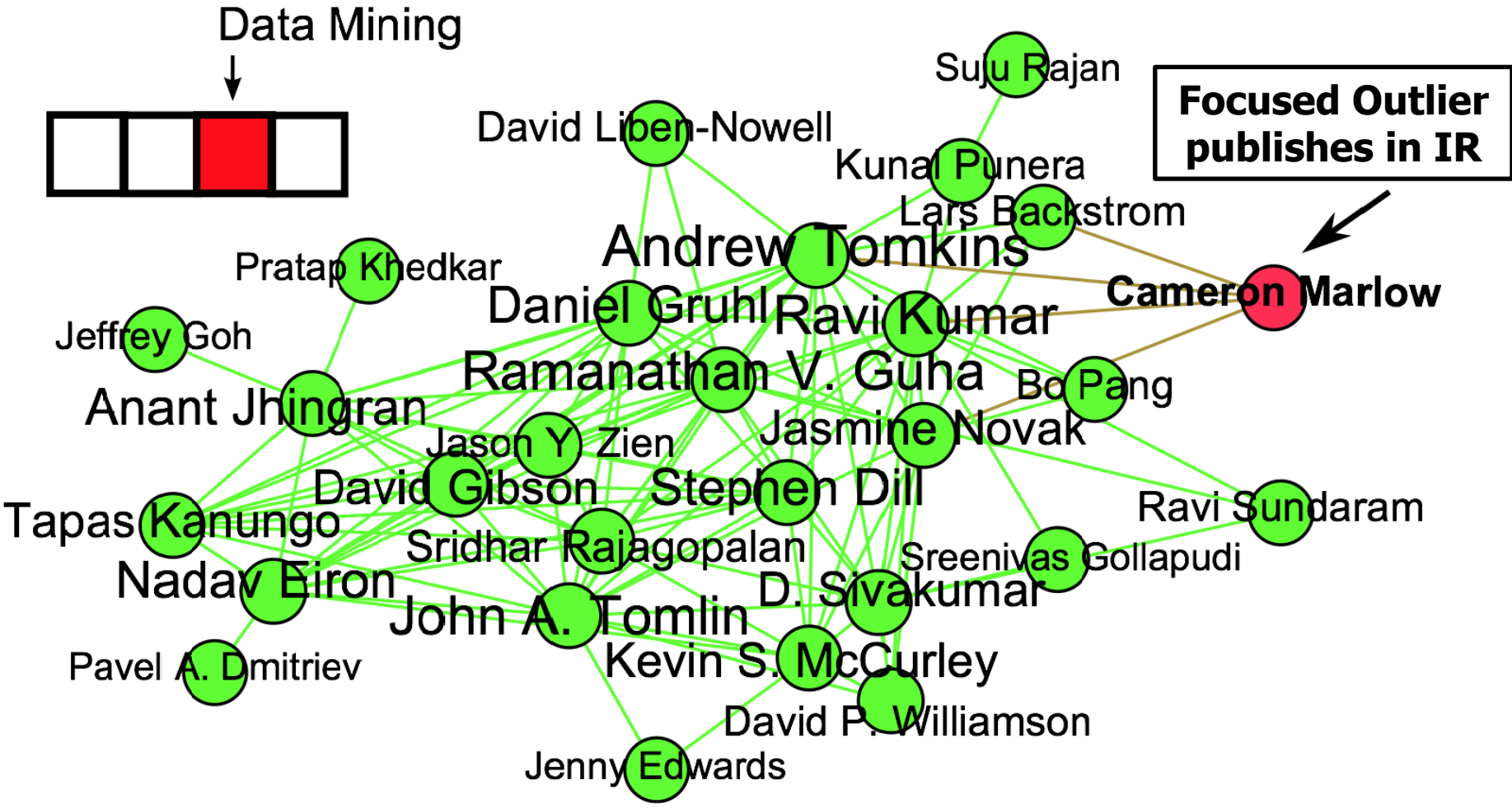
(b) precision

deflated focus attributes increased (easier) from left to right

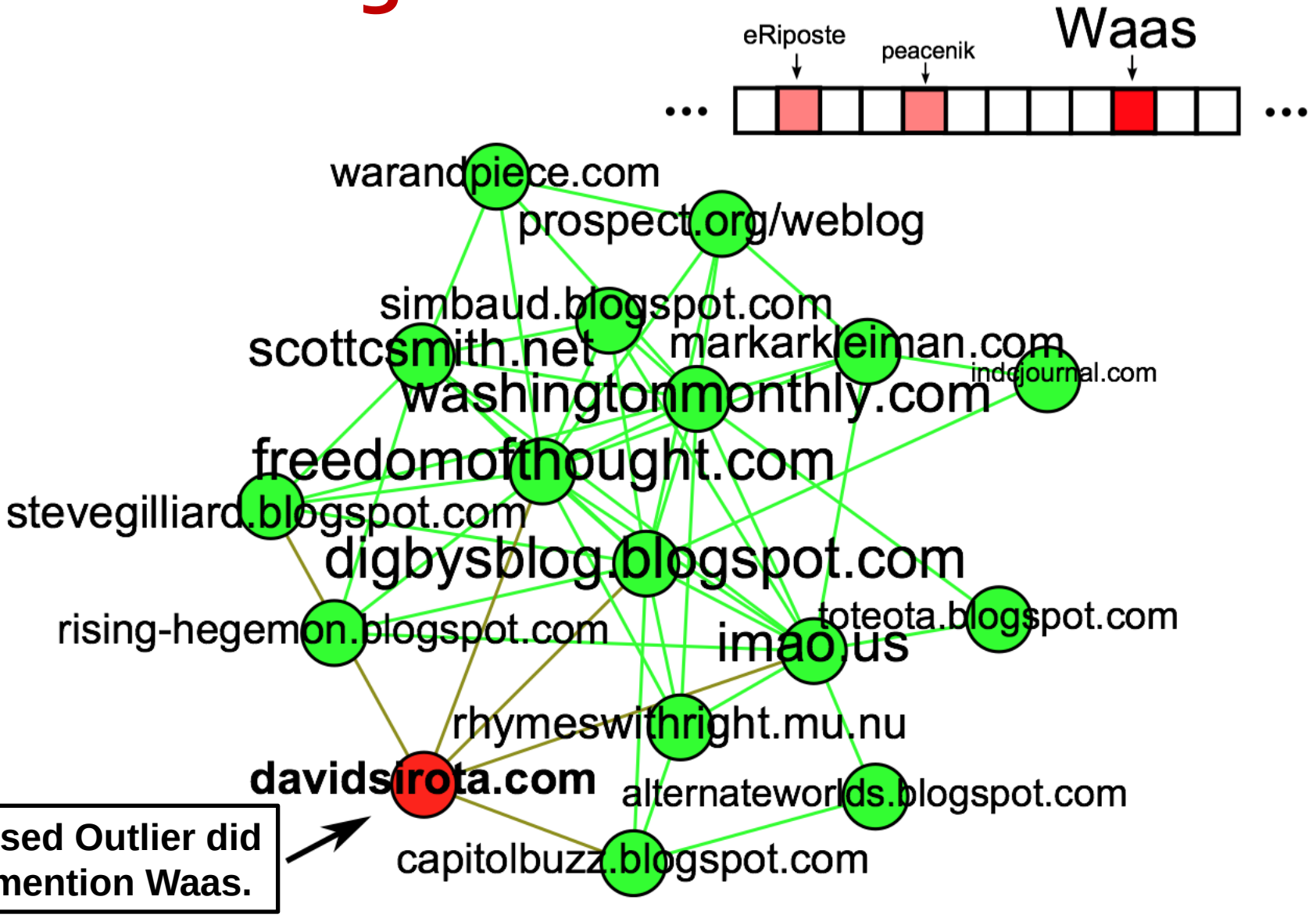
Disney: Amazon co-purchase graph



DBLP co-authorship graph



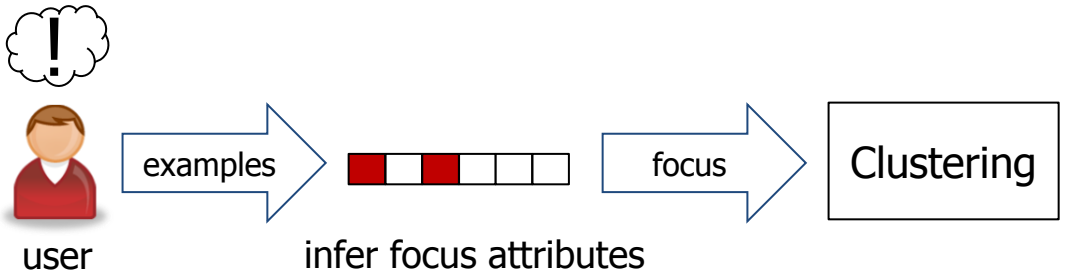
Political blogs citation graph



Summary

A new graph mining paradigm
where the **focus** steers graph mining
according to user preference.

A new problem formulation
Focused Clustering & **O**utlier detection



Thanks! Any questions?

Bryan Perozzi (bperozzi@cs.stonybrook.edu)