Graph-Based Anomaly Detection:

Problems, Algorithms and Applications

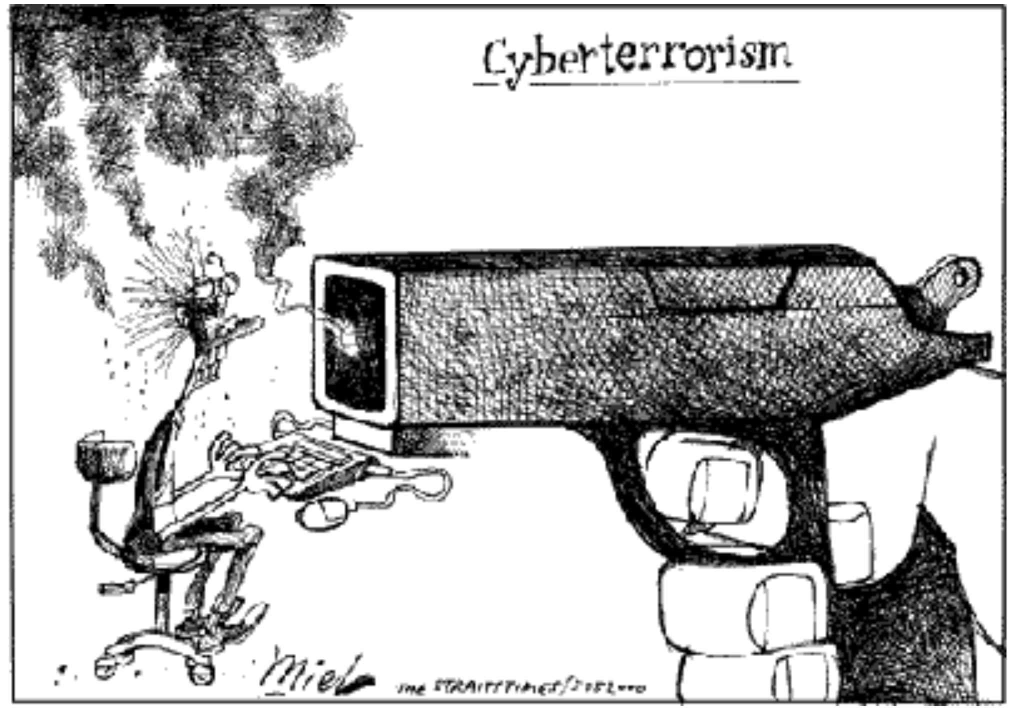
Leman Akoglu

Carnegie Mellon University
Heinz College

Anomaly: That stands out

https://en.wikipedia.org/wiki/August_Landmesser

Anomaly Detection: Many Use-cases

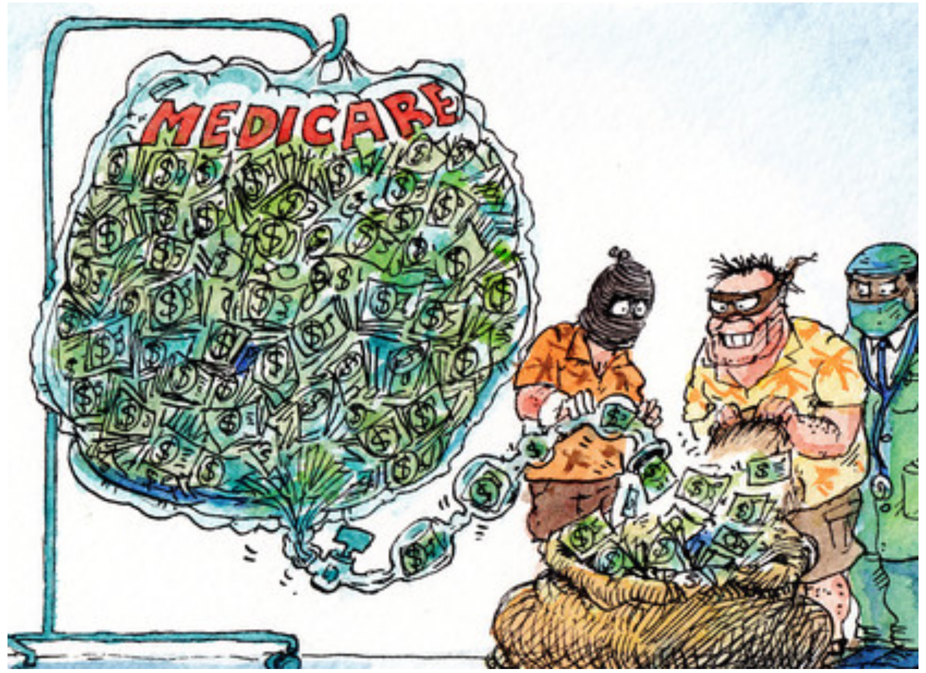


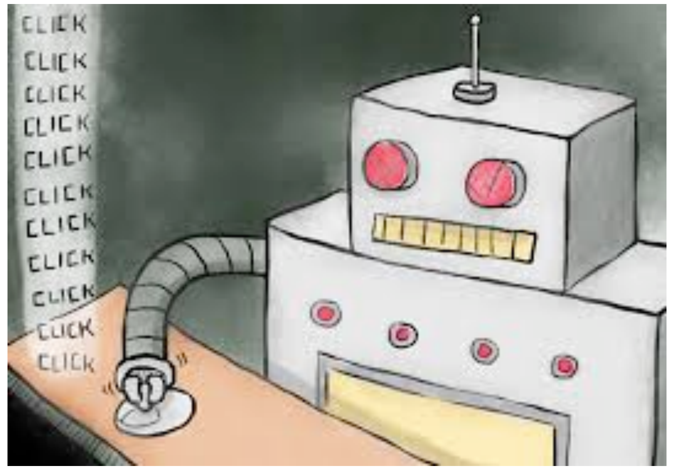
amazon

tripadvisor

yelp

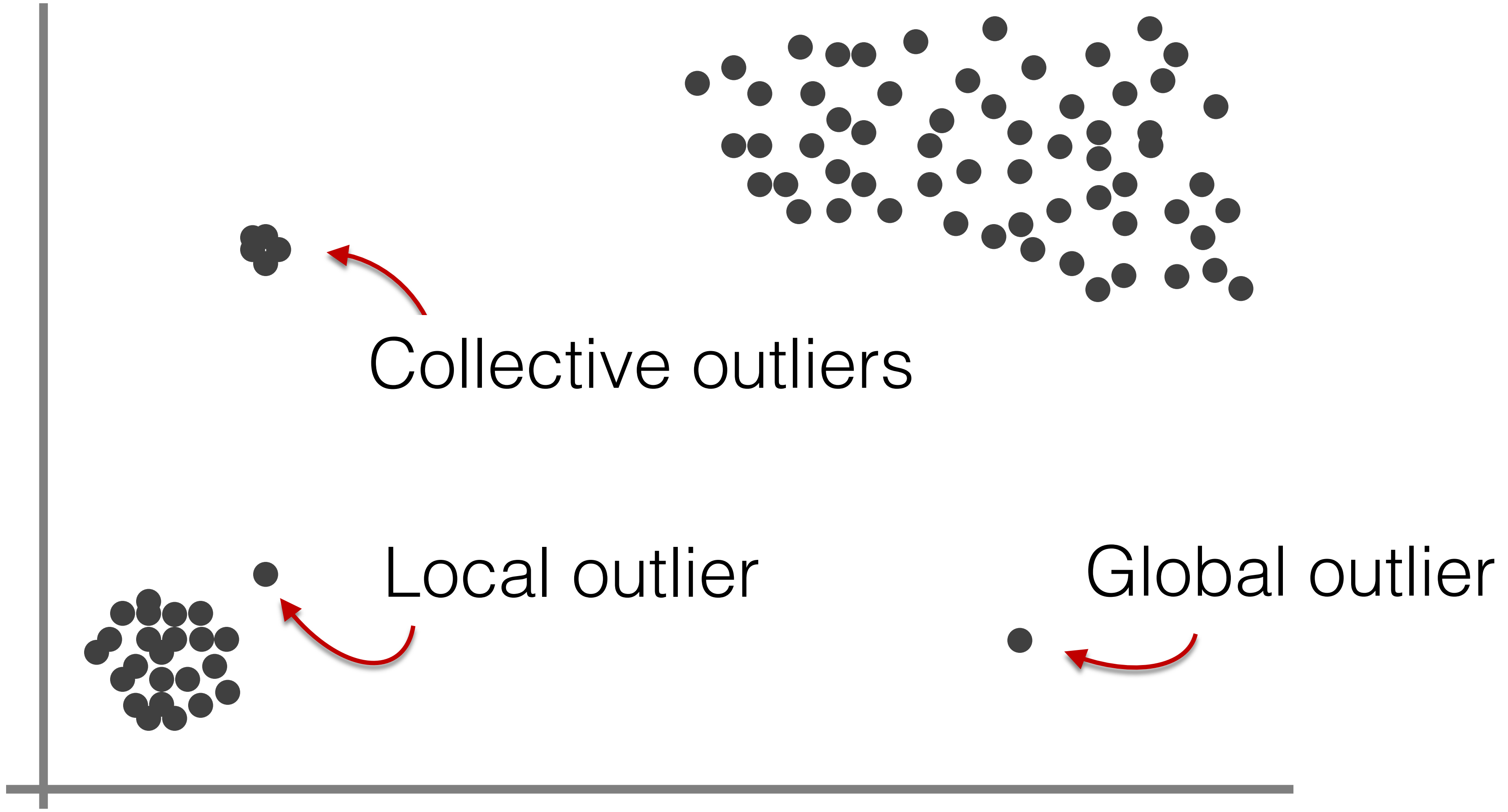
Expedia





Formalizing Anomaly Detection

- Concrete problem settings exist. e.g./esp. for point-cloud data



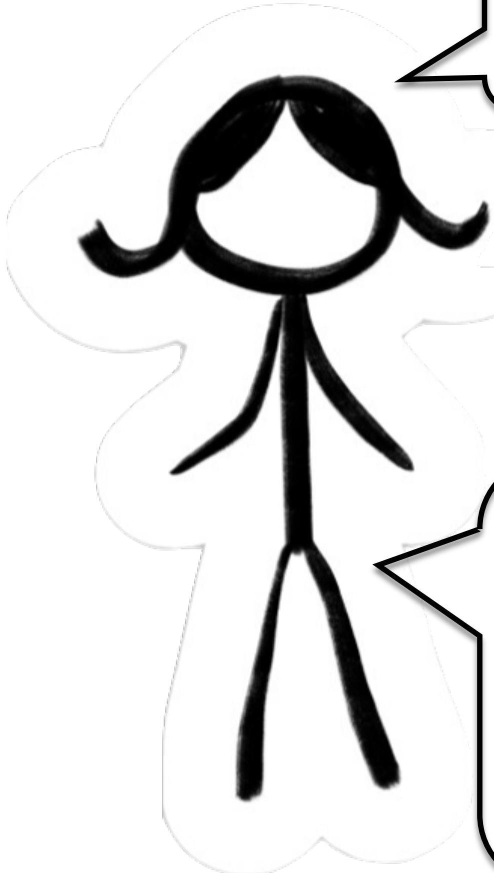
- Real-world… A bit more complex.

Formalizing Anomaly Detection

Given <DATA>, Find <ANOMALIES>

e.g. (accounting) Given millions of transactions, find abnormalities

We heard you work on anomaly detection.



Yes, I am very excited. Tell me more.

We have lots of data, and want to find anomalies.

OK, wait, tell me what your REAL PROBLEMS are.

Why do you want to detect anomalies?

What do you consider to be an anomaly?

Formalizing Anomaly Detection

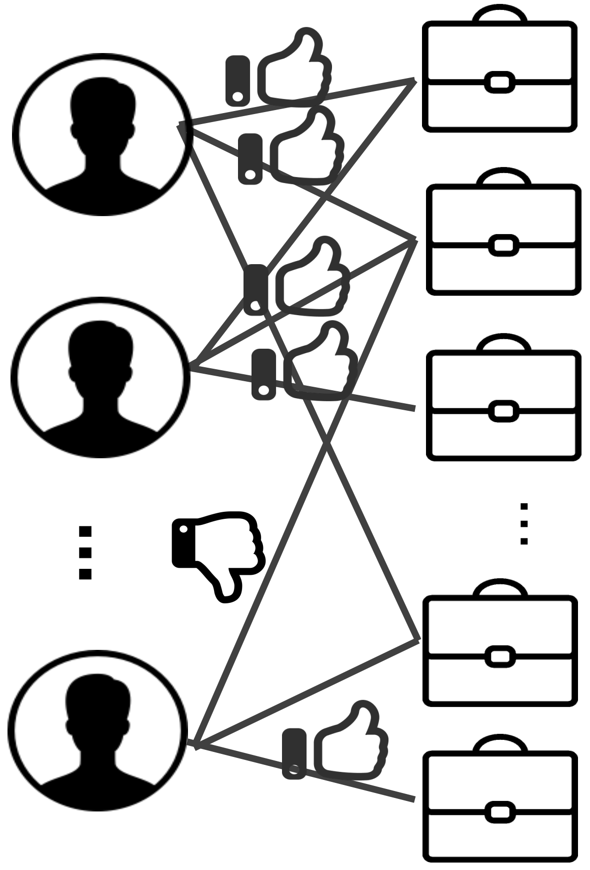
Tell me what your REAL PROBLEMS are.

We want to find errors, inefficiencies, malfeasance… We want to save $$$. We also want YOU to find all unknown anomalies.

Hmm… OK…?

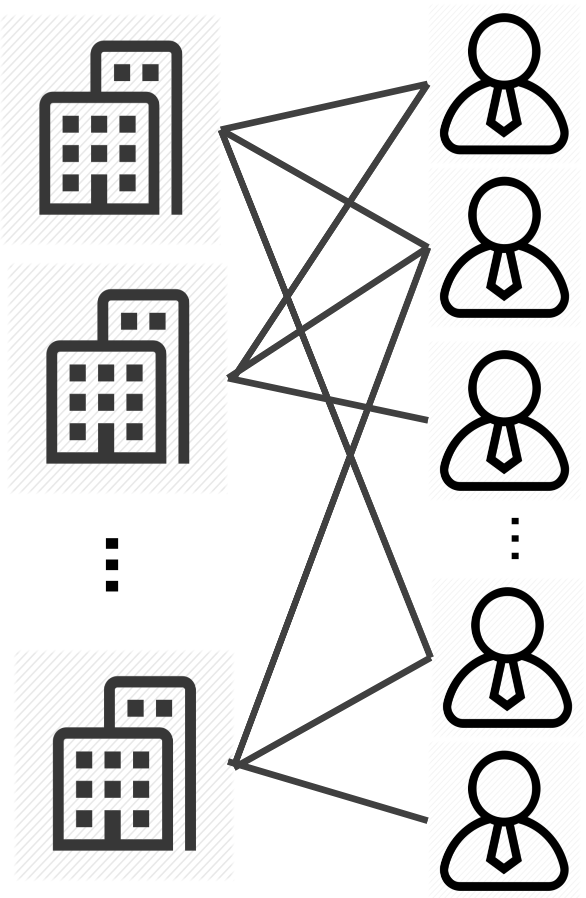
Graph-based Anomaly Detection

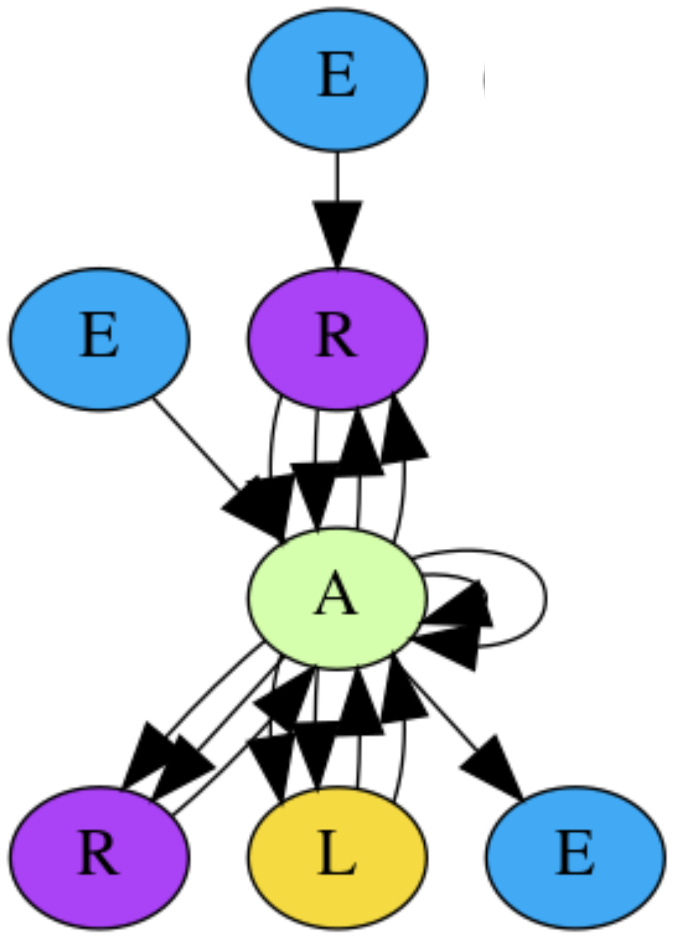
- Often, underlying data is unmistakably relational:



user-business reviews

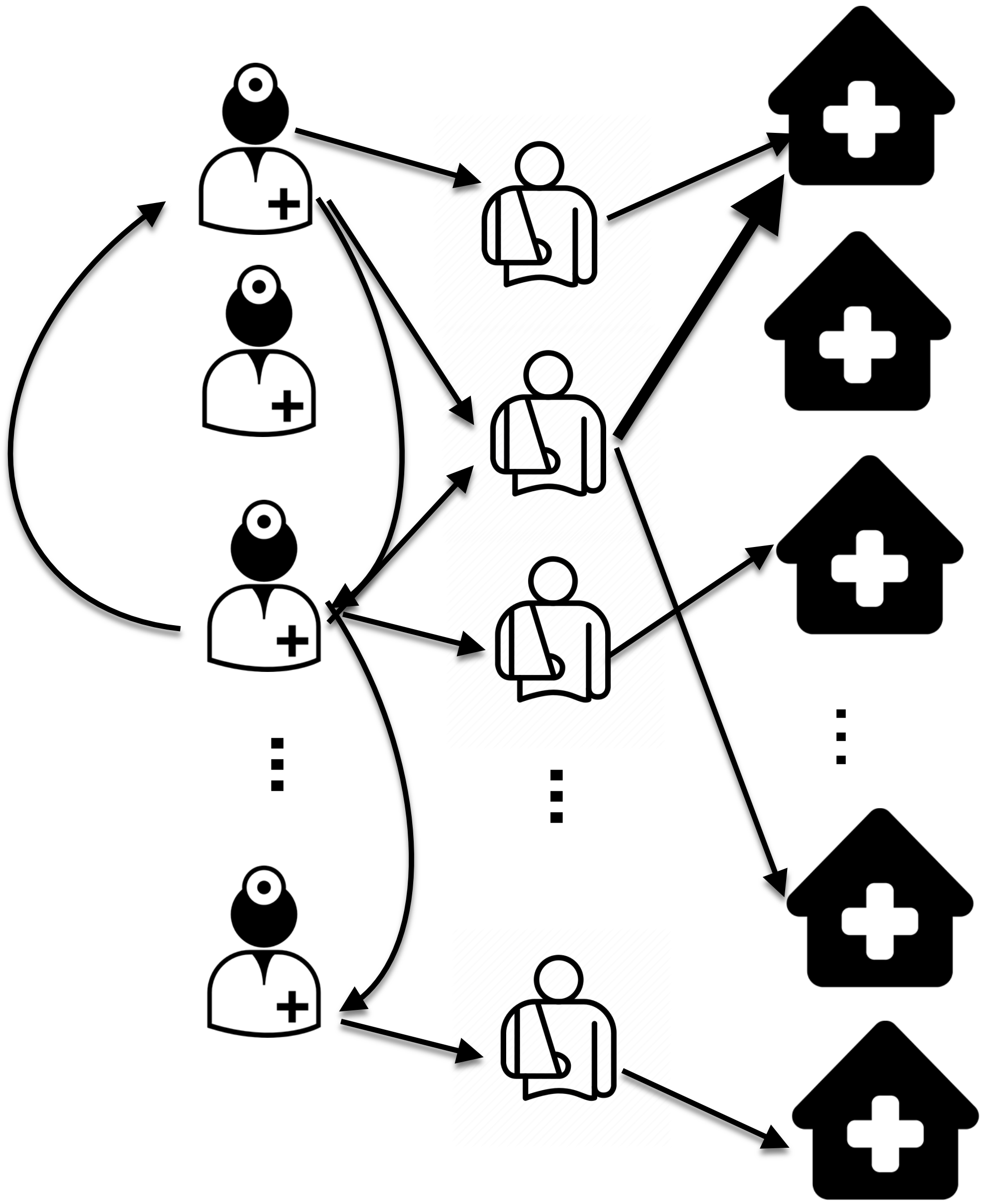
employer-employee





account transactions

physician-patient-provider



Graph-based Anomaly Detection

Several surveys and tutorials:

[Survey] Graph-based Anomaly Detection and Description: A Survey. [Akoglu+]
Data Mining and Knowledge Discovery (DAMI), May 2015.

[Tutorial] Fraud Detection through Graph-Based User Behavior Modeling. [Beutel+]
ACM CCS 2015.

[Tutorial] Social Media Anomaly Detection: Challenges and Solutions. [Liu & Chawla]
ACM SIGKDD 2015.

[Survey] False Information on Web and Social Media: A Survey. [Kumar & Shah]
arXiv:1804.08559

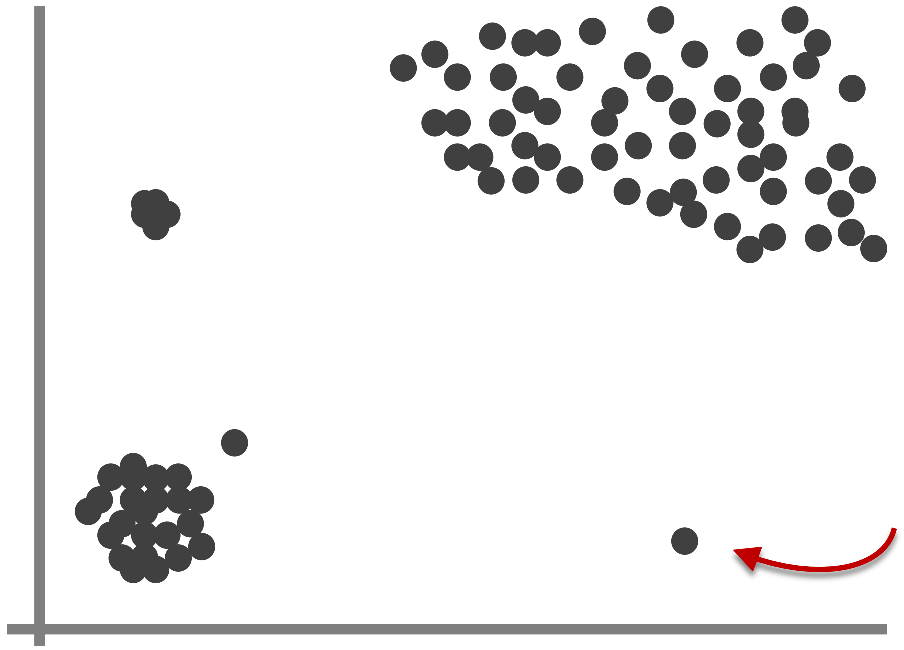
Challenges

Problem: Given <Data>, Find <Anomalies> s.t. <Constraints>

1. <Data> : Graph heterogeneity (node/edge labels, attributes, multi-edges, edge weights, edge timestamps, etc.)
   - How or whether to "fold" meta-data into a graph
2. <Anomalies> : Definition/Formalization of anomalies (e.g., group anomalies vs. anomalous groups)
   - Heterogeneity exacerbates the issue
3. <Constraints> : System/Application requirements e.g., distributed/streaming/massive data, attribution (who), explainability (why)

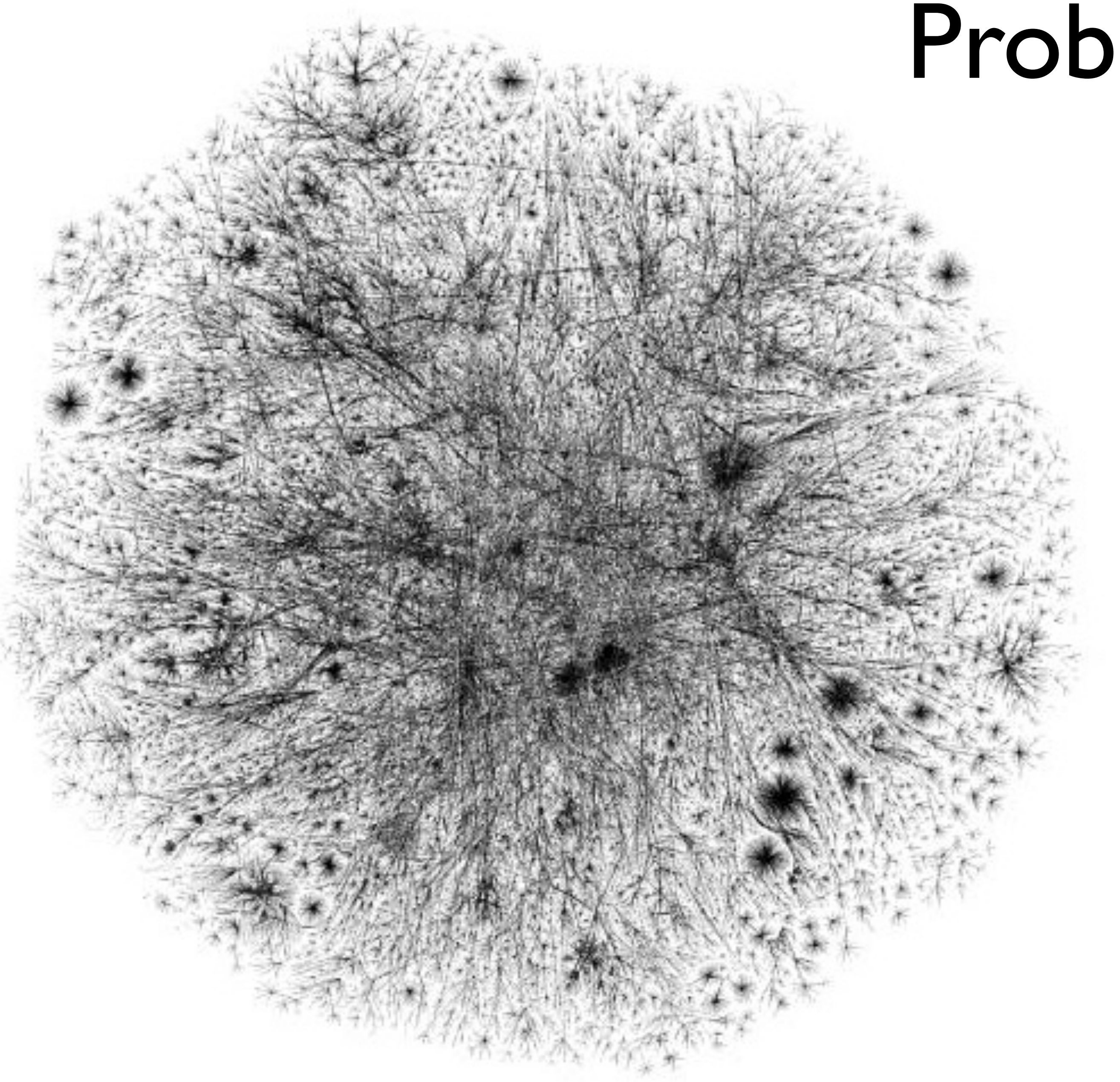
Outline

- Anomaly Detection: Motivation, Formalism, Challenges
- Graph-based Anomaly Detection
  - General-purpose (single graph)
    - Global – anomalous nodes
    - Local – group anomalies
    - Collective – anomalous groups
  - Specialized (graph database)
- Recent Trend: Deep Anomaly Detection



Anomalous nodes (global)

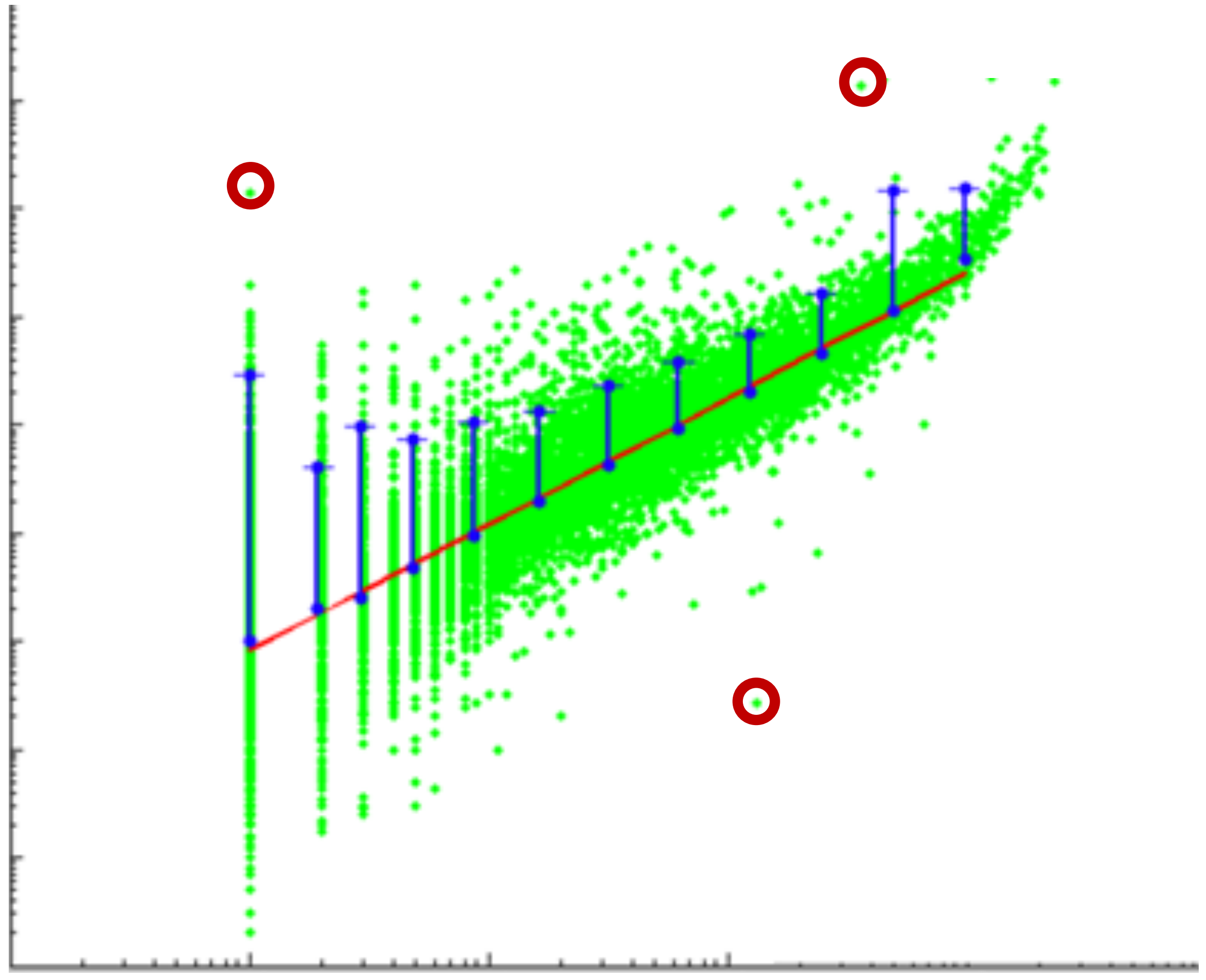
Problem Sketch:



plain, weighted, directed

?

?

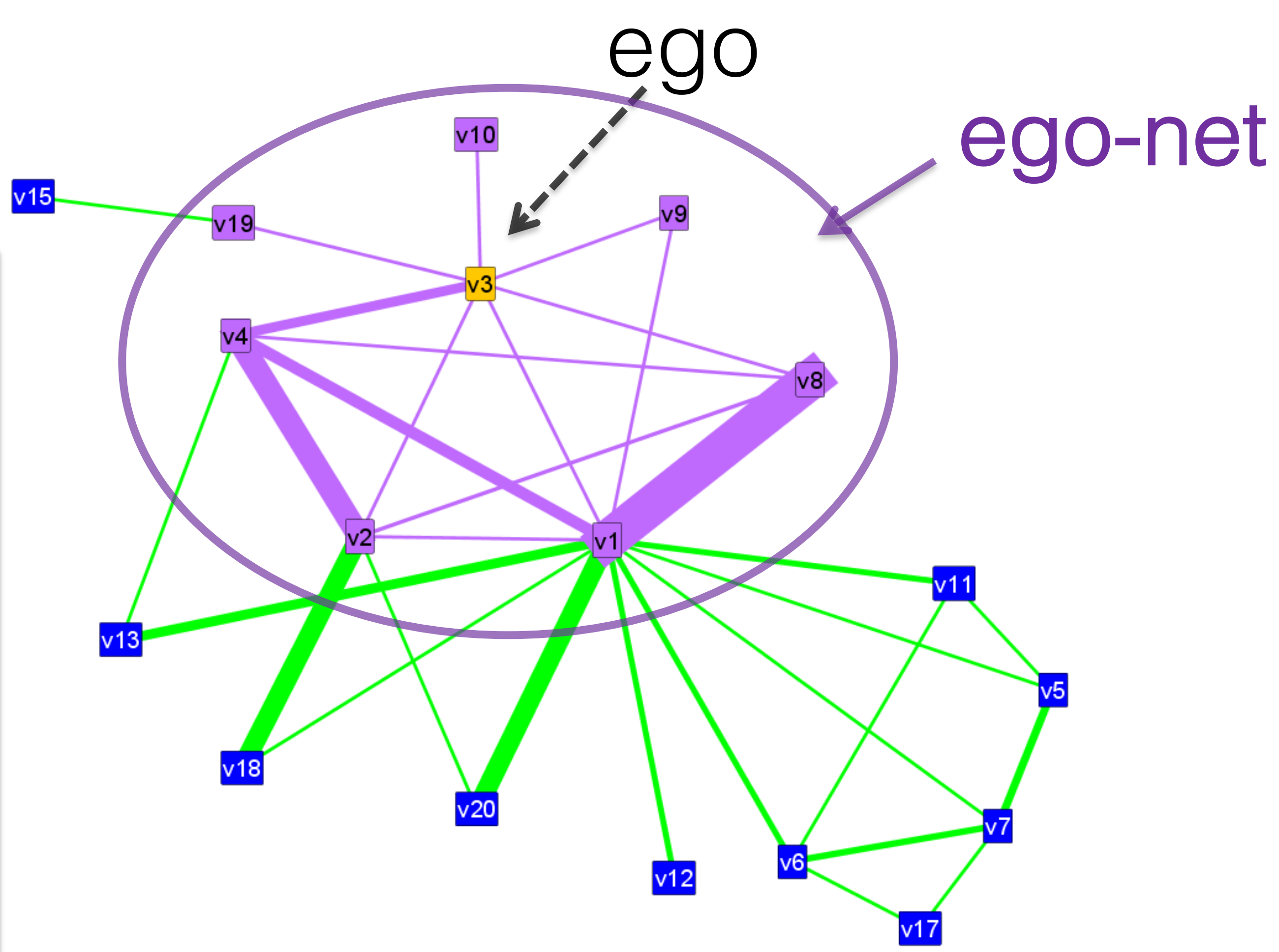


?

Anomalous nodes (global): OddBall

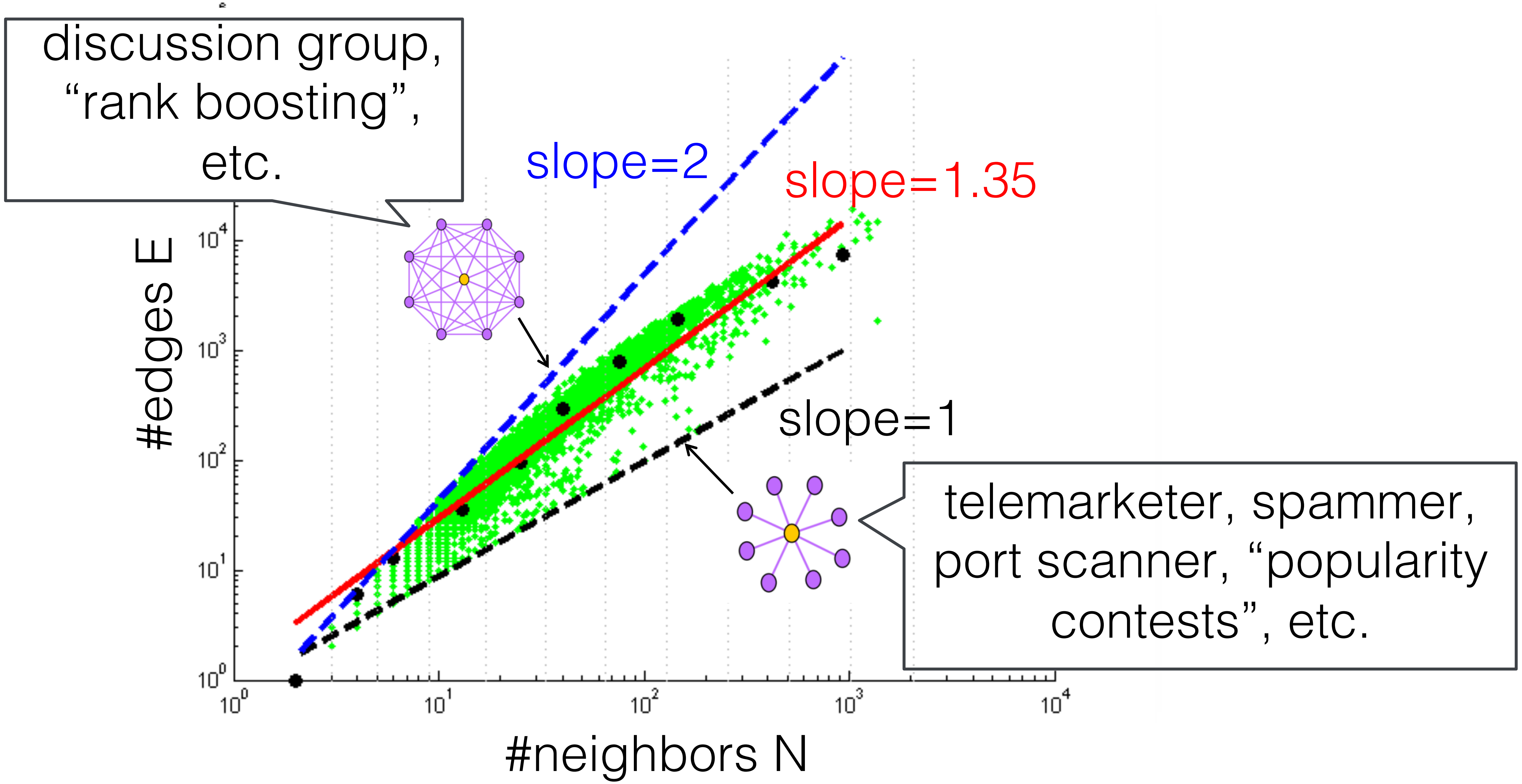
Problem Setting:

- For each node
  - Extract ego-net (1-hop neighborhood)
  - Extract ego-net features
- Find patterns ("laws")
- Detect outliers (distance to patterns)

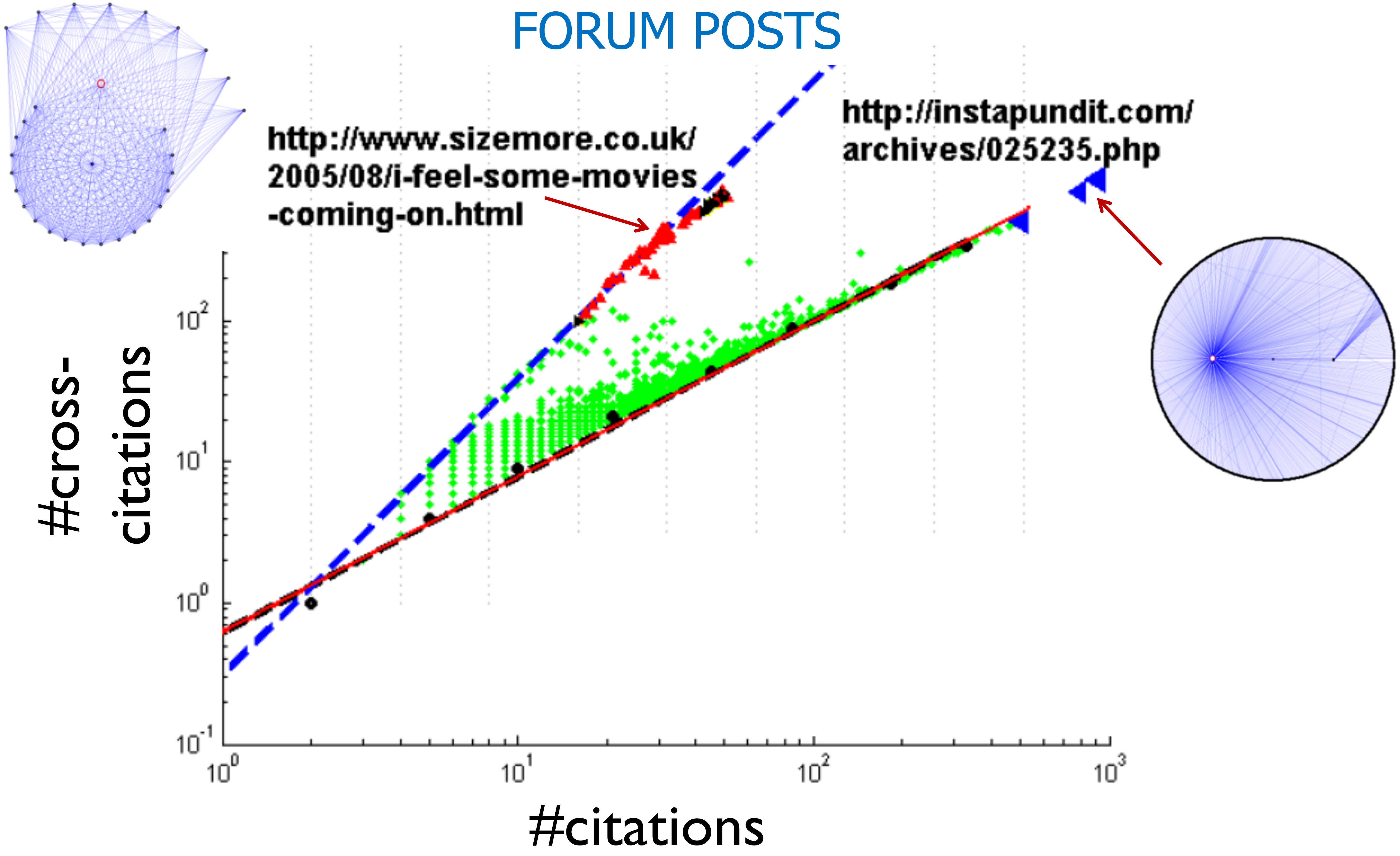


OddBall: Spotting Anomalies in Weighted Graphs. [Akoglu+] PAKDD 2010.

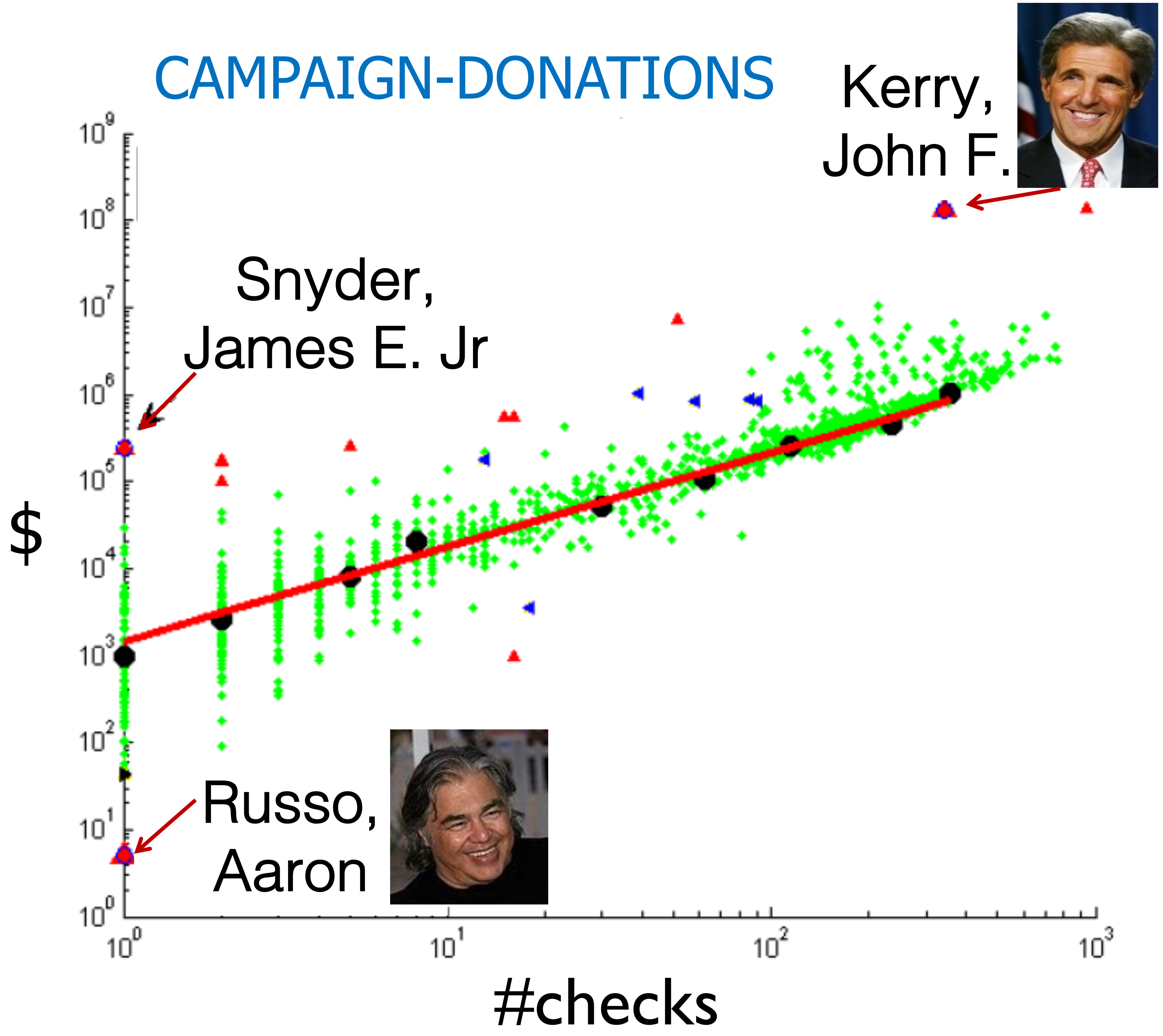
Anomalous nodes (global): OddBall



Anomalous nodes (global): OddBall

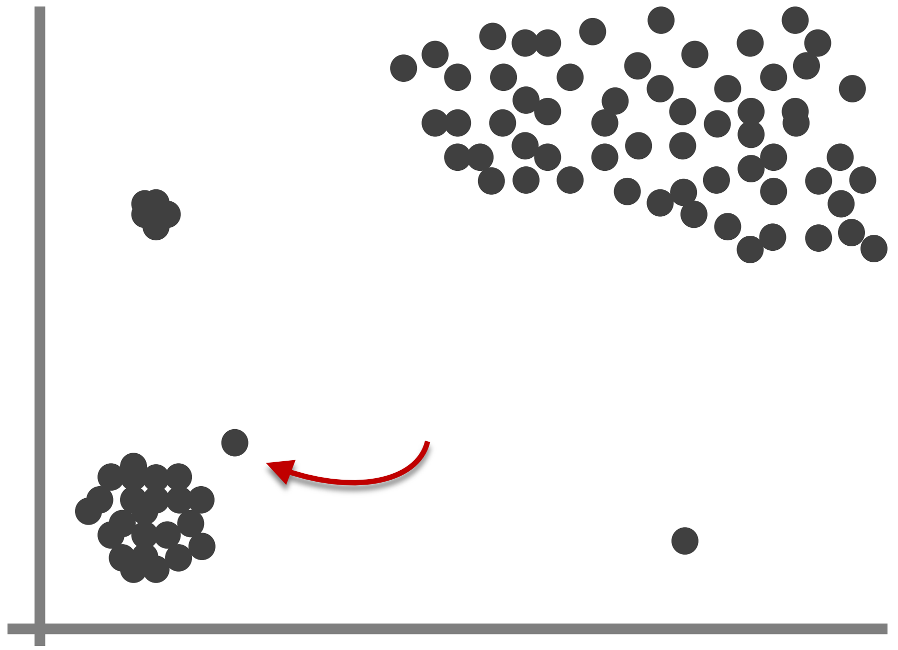


Anomalous nodes (global): OddBall



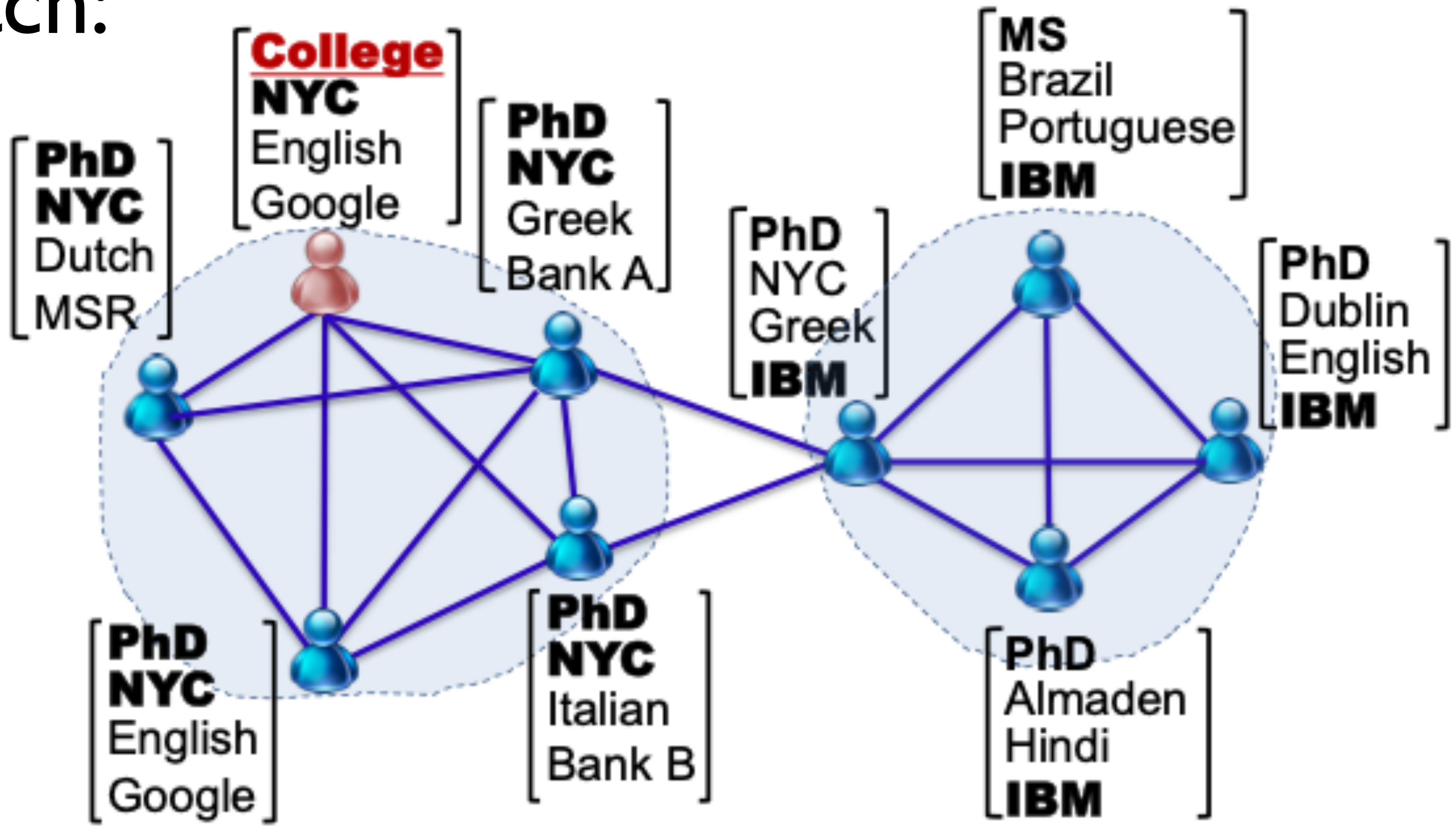
Outline

- Anomaly Detection: Motivation, Formalism, Challenges
- Graph-based Anomaly Detection
  - General-purpose (single graph)
    - Global – anomalous nodes
    - Local – group anomalies
    - Collective – anomalous groups
  - Specialized (graph database)
- Recent Trend: Deep Anomaly Detection



Group anomalies (local)

Problem Sketch:



(left) community "focuses" on {`degree`, `location`} (right) "focuses" on `work`.

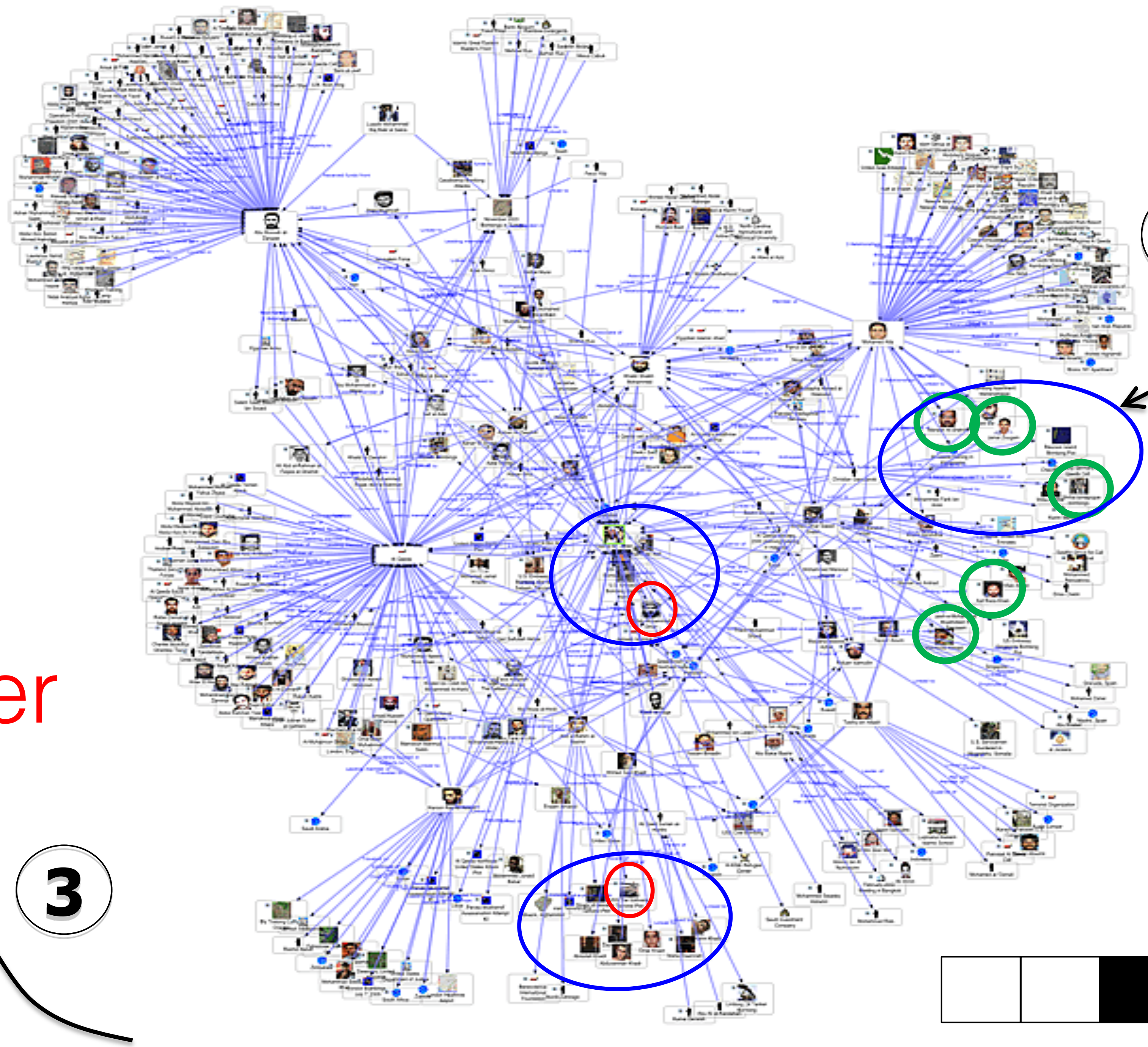
L. Akoglu

Carnegie Mellon

Group anomalies (local): FocusCO

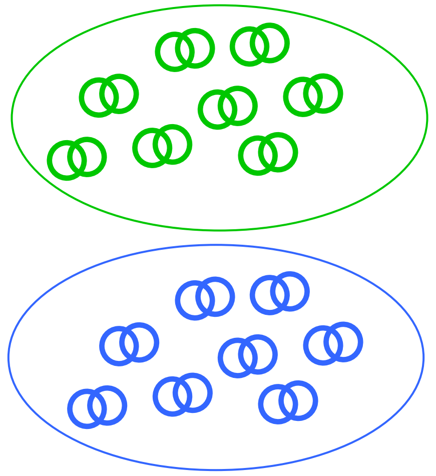
cluster
extraction &
(cluster) outlier
detection

3



1

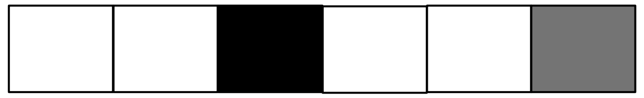
similar pairs



dissimilar pairs

2

“**focus**”
estimation

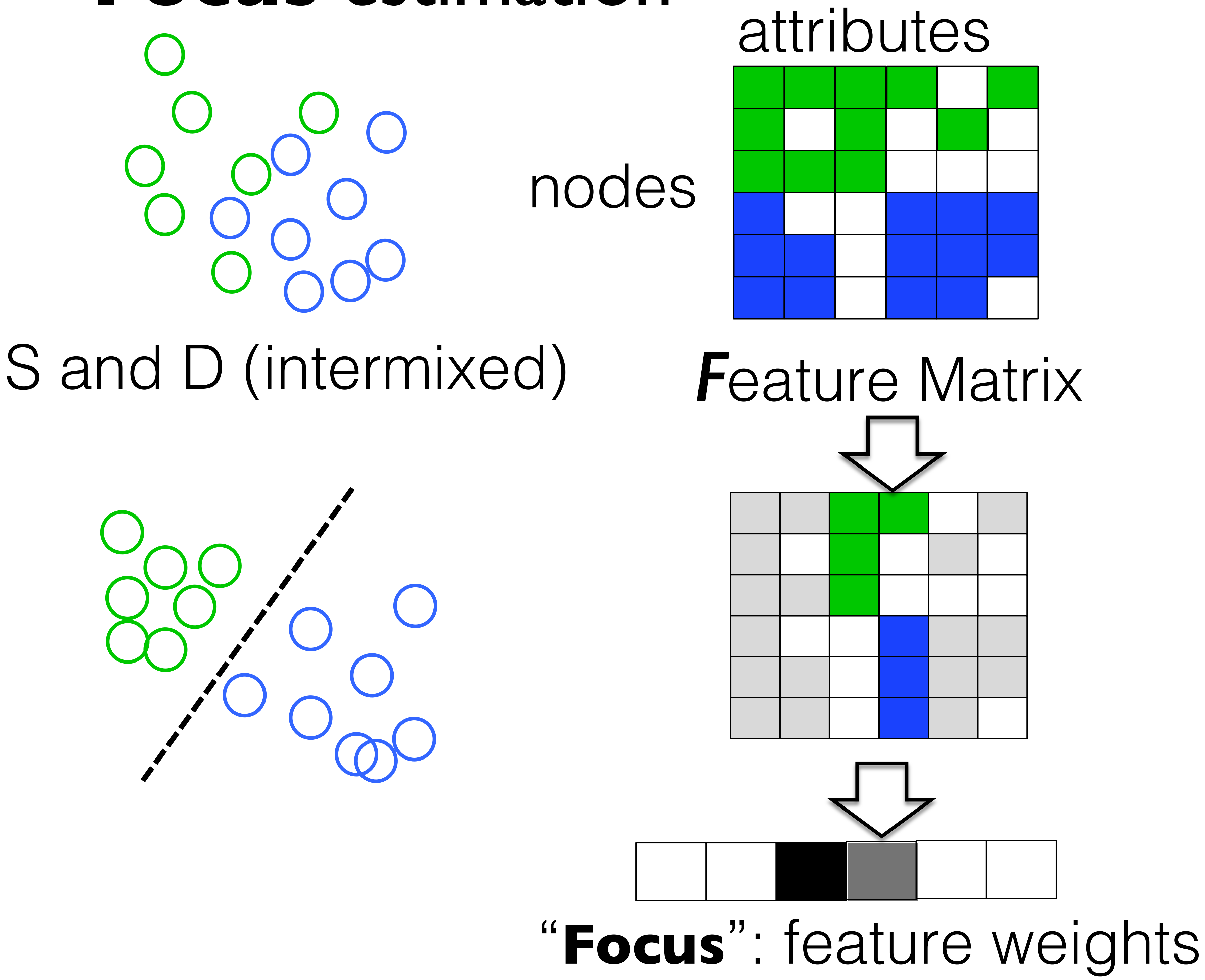


age

gender

Group anomalies (local): FocusCO

- **Focus** estimation

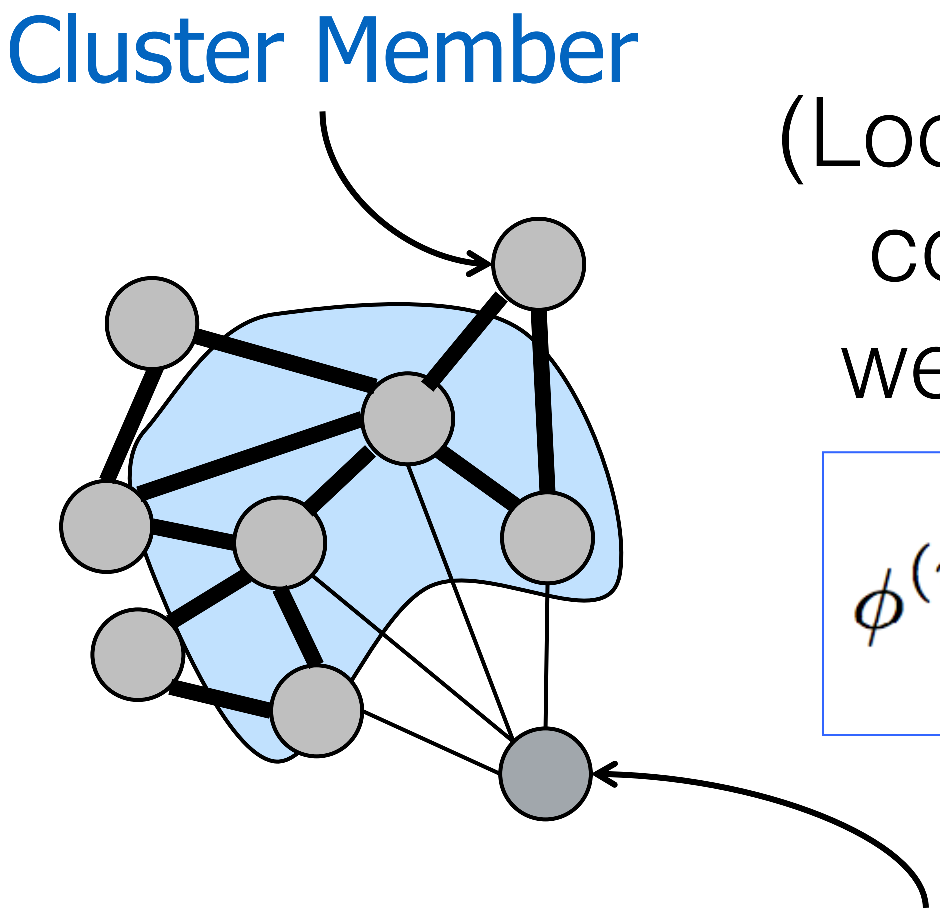


- Clusters & **Local** outliers

Group anomalies (local): FocusCO

- **Focus** estimation

- Clusters & **Local** outliers

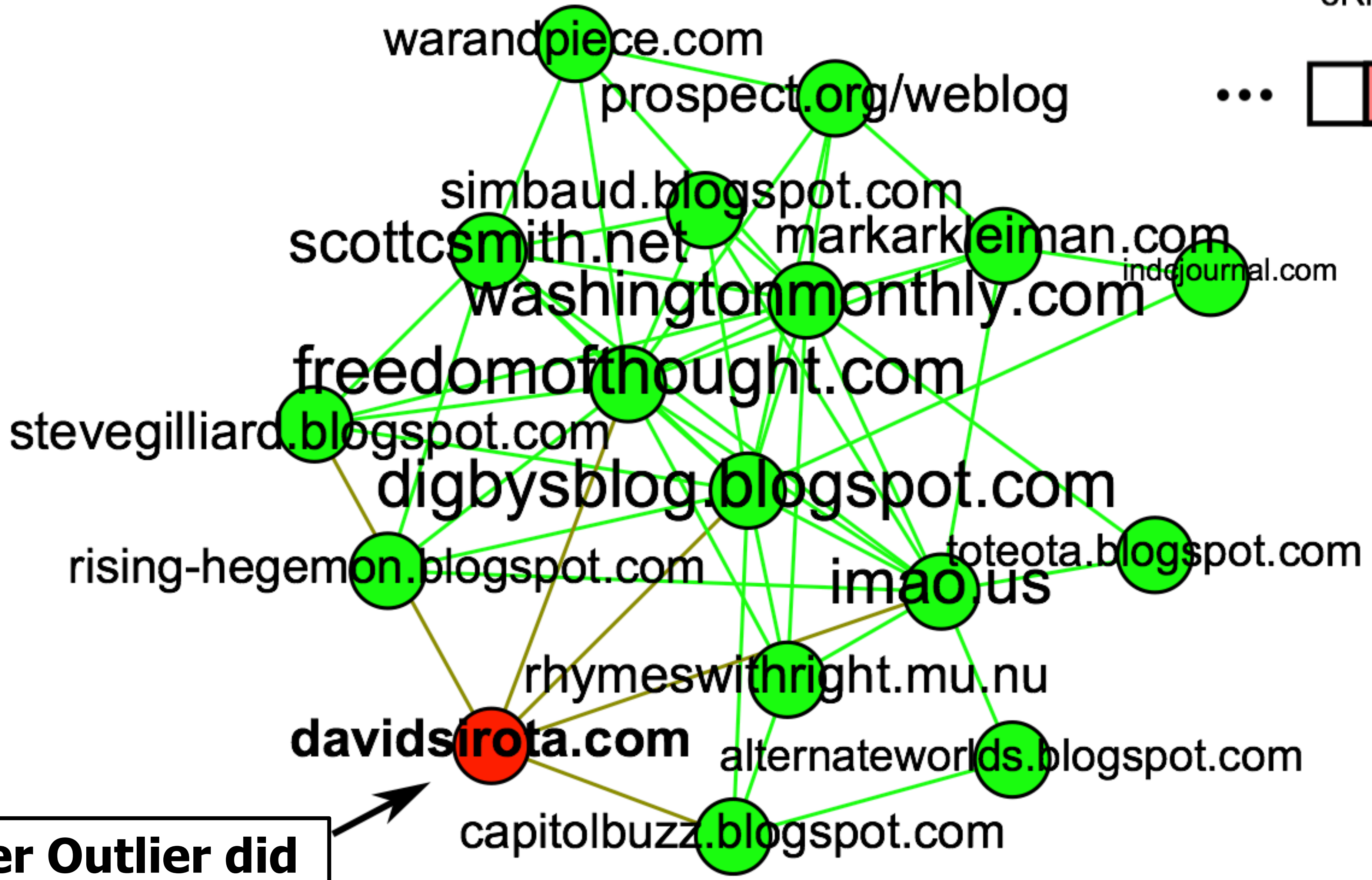


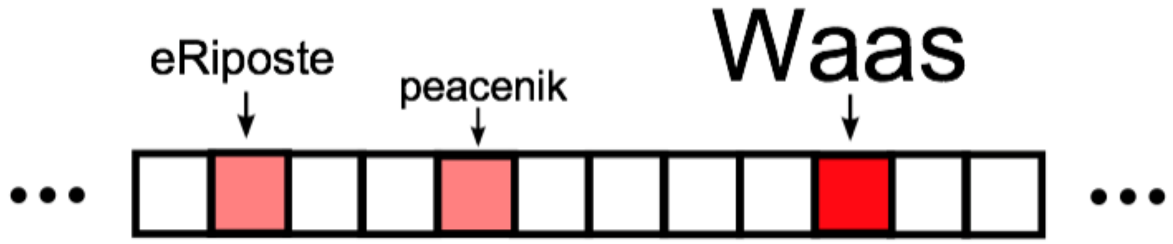
(Local) clustering obj.:
conductance $\phi^{(w)}$
weighted by **focus**

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

Focused Outlier
node with
many (but) weak ties

Group anomalies (local): FocusCO



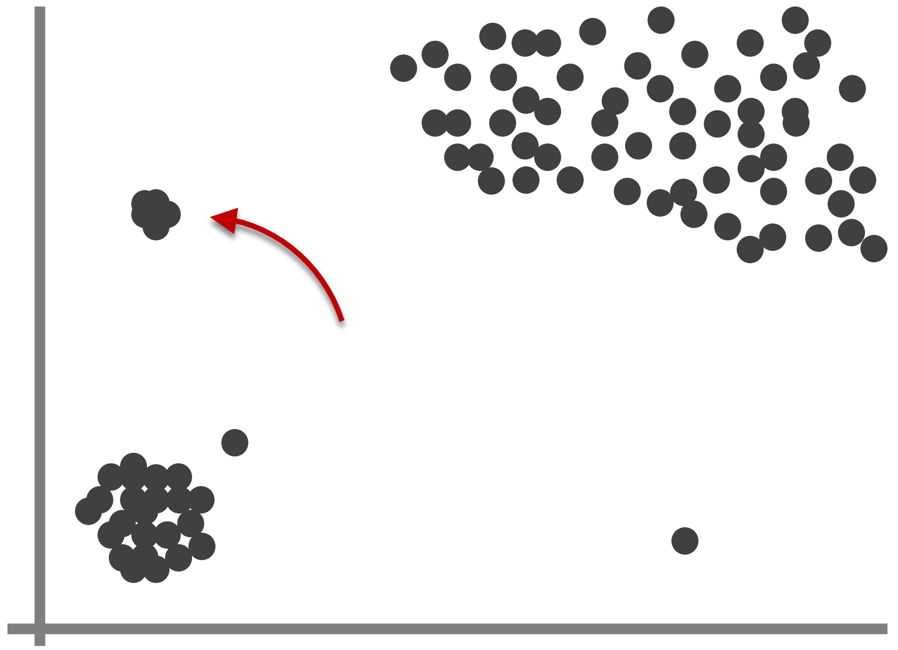


Cluster Outlier did not mention Waas.

Liberal Cluster in Political Blogs Graph

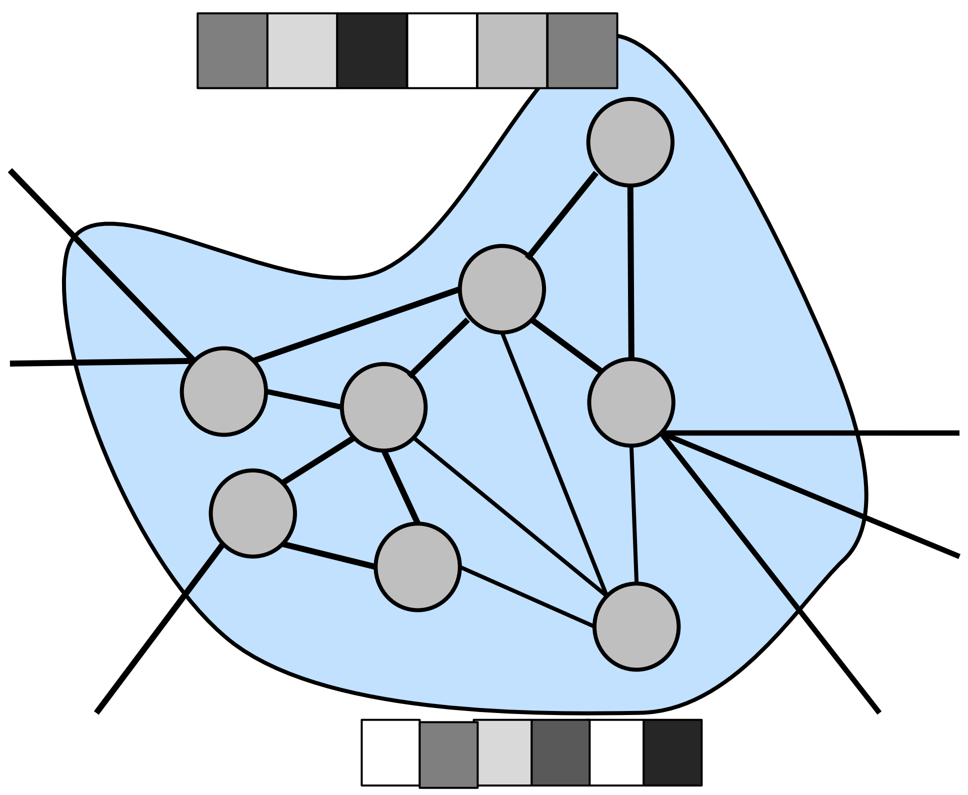
Outline

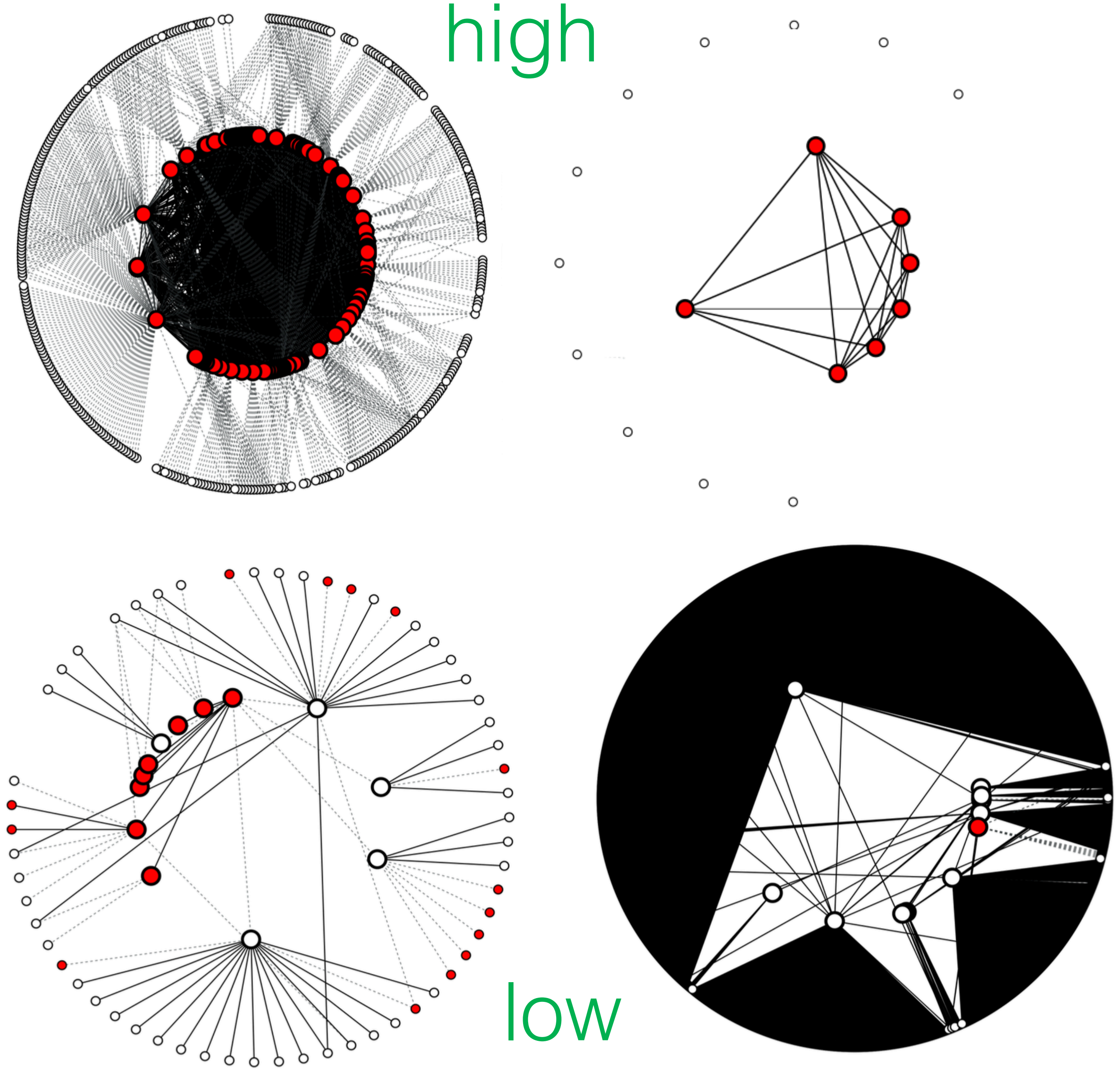
- Anomaly Detection: Motivation, Formalism, Challenges
- Graph-based Anomaly Detection
  - General-purpose (single graph)
    - Global – anomalous nodes
    - Local – group anomalies
    - Collective – anomalous groups
  - Specialized (graph database)
- Recent Trend: Deep Anomaly Detection

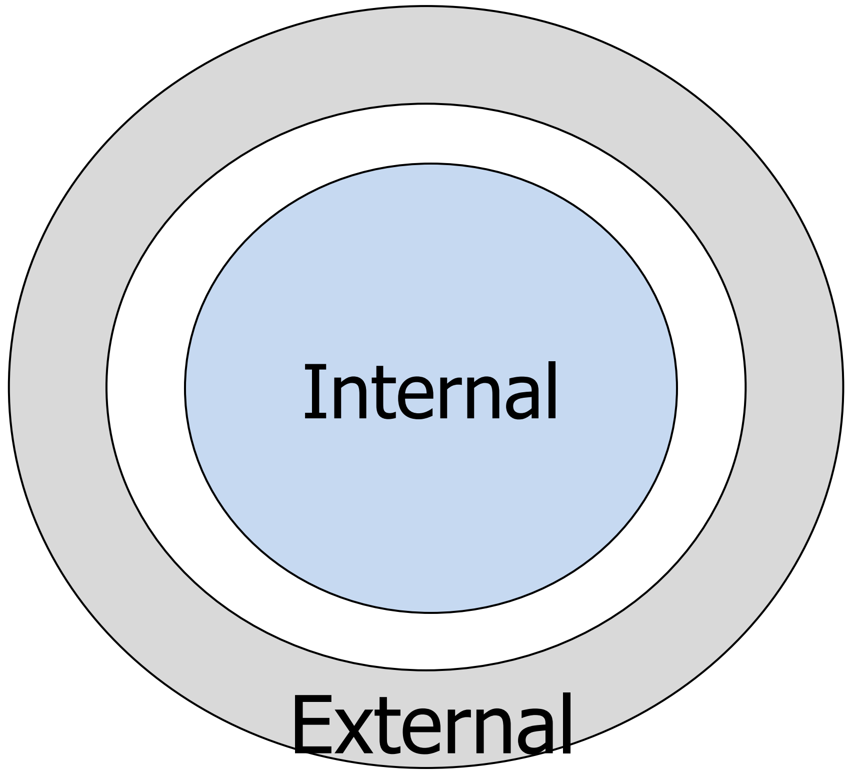


Anomalous groups (collective)

- Problem Sketch:
  Given a node-attributed subgraph, how to define "normality"?



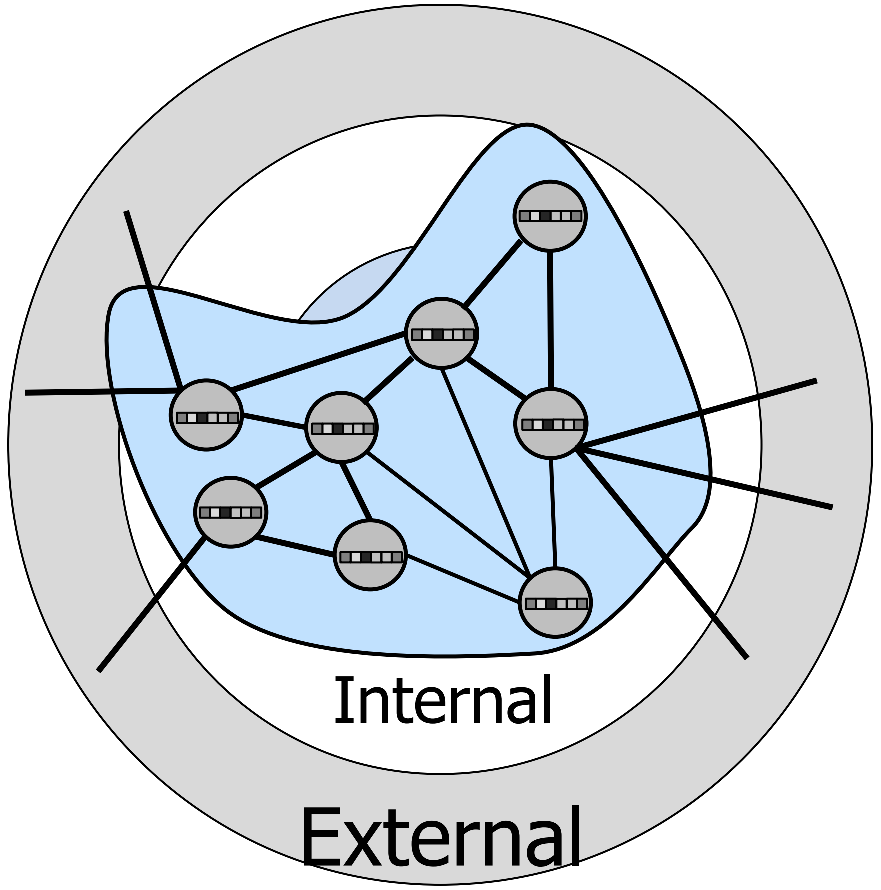




Anomalous groups (collective): AMEN

Problem Setting:

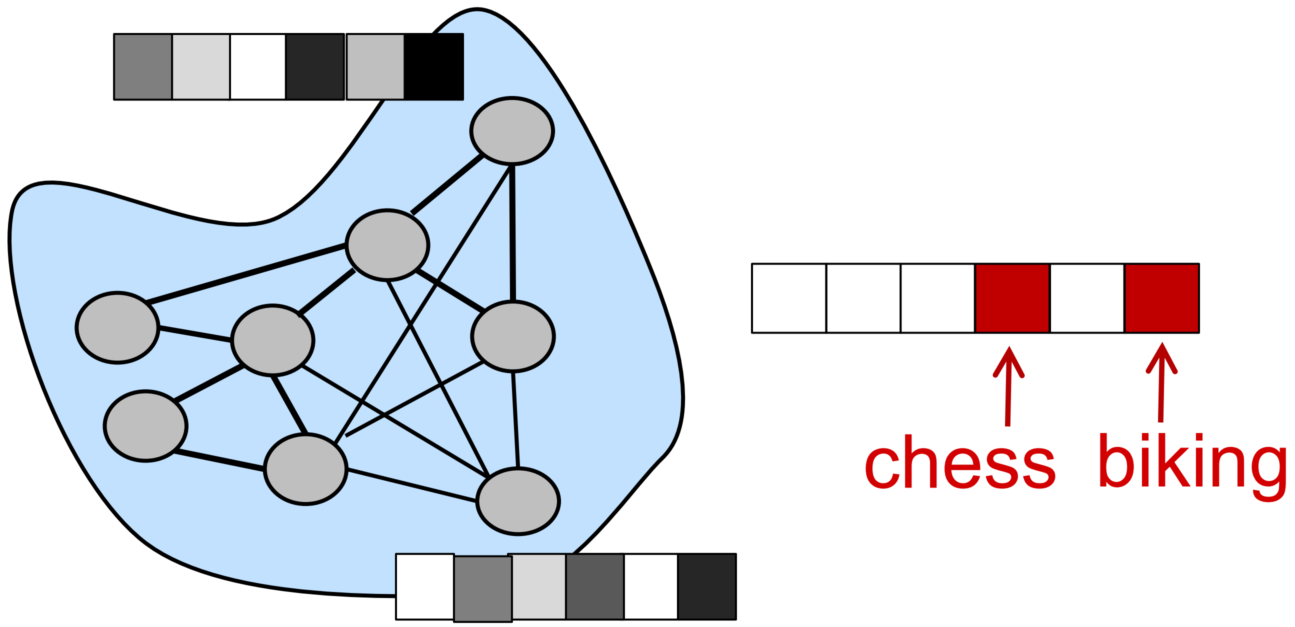
- Given a subgraph in a node-attributed graph
- Identify (subgraph) "**focus**" such that
  - Internal nodes are
    structurally dense & coherent in **focus**
  - External nodes are structurally sparse
    or not-surprising, or different in **focus**

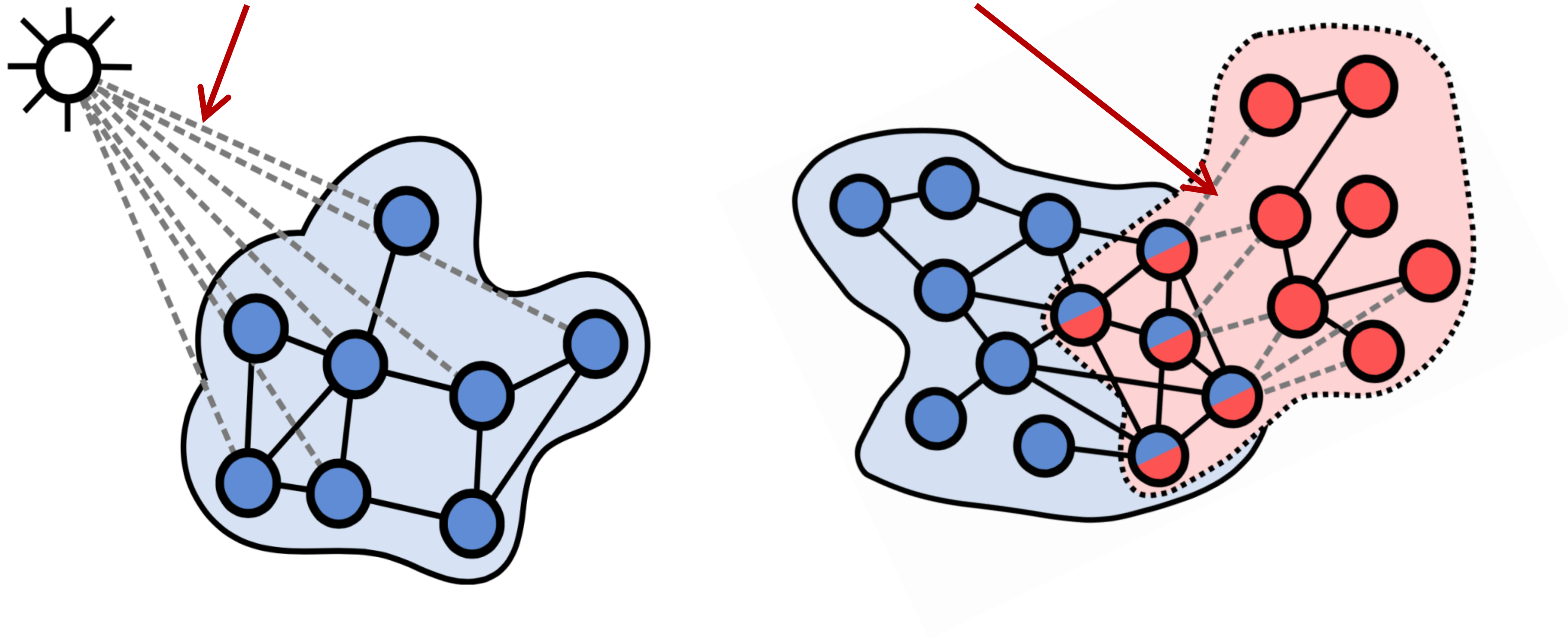


Scalable Anomaly Ranking of Attributed Neighborhoods. [Perozzi & Akoglu] SIAM SDM, 2016.

Anomalous groups (collective): AMEN

- Internal nodes are structurally dense & coherent in **focus**
- External nodes are structurally sparse or **not-surprising**, or different in **focus**





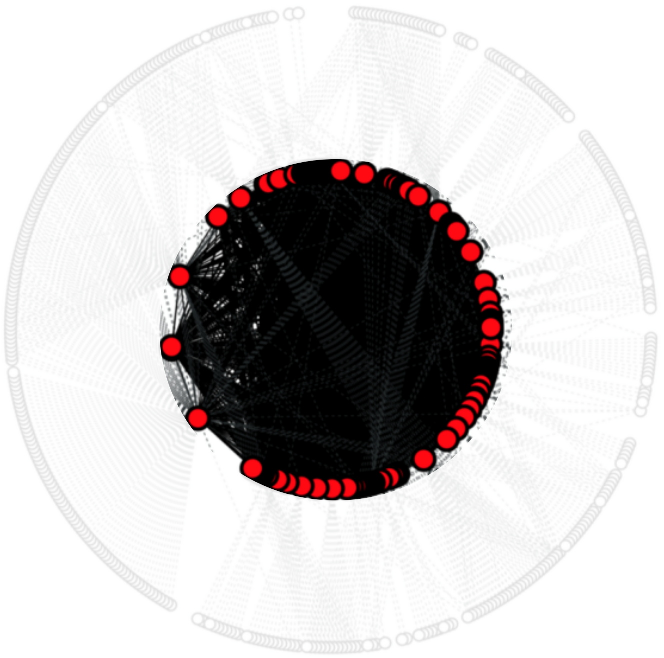
Anomalous groups (collective): AMEN

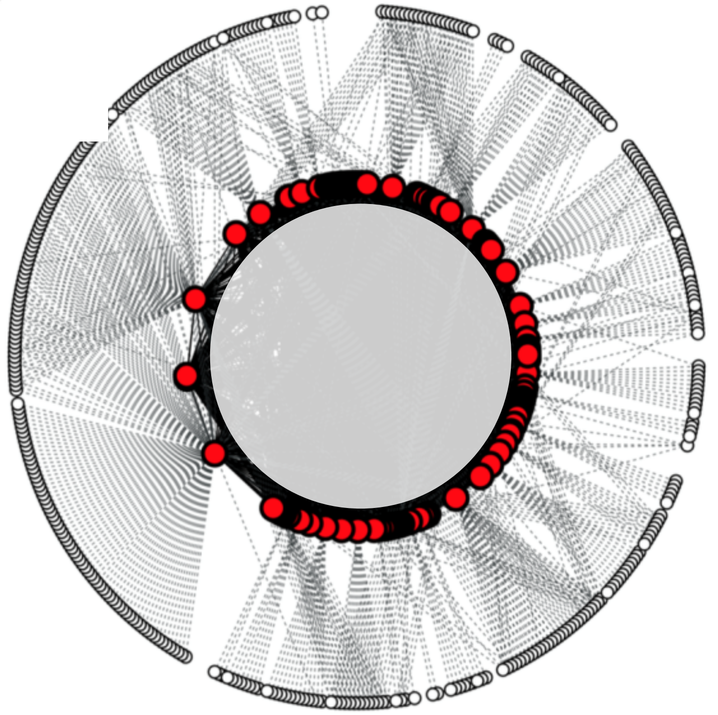
- Measure of Normality:

null model similarity "focus" vector

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

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





Anomalous groups (collective): AMEN

- Estimating Normality:

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

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

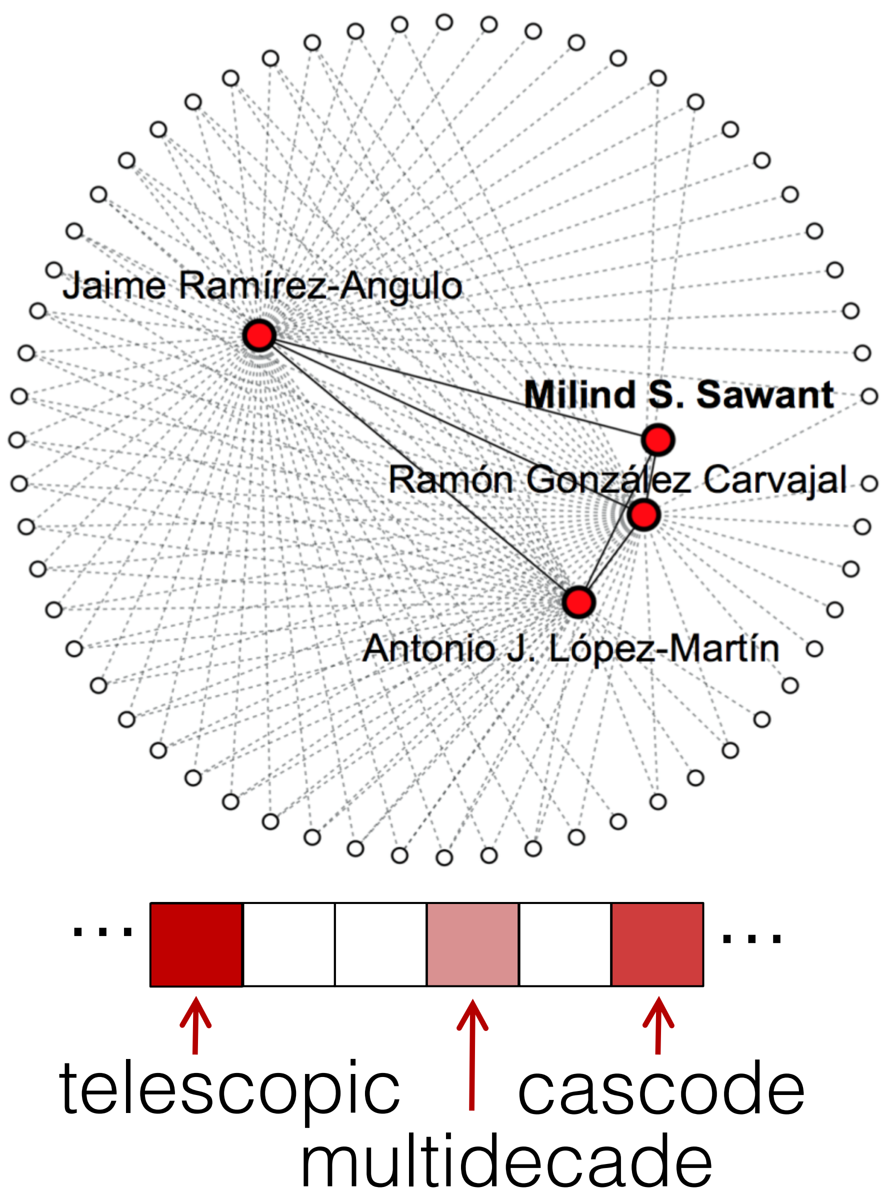
$$\max_{\mathbf{w_C}} \quad \mathbf{w_C}^T \cdot (\hat{\mathbf{x}}_I + \hat{\mathbf{x}}_E)$$

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

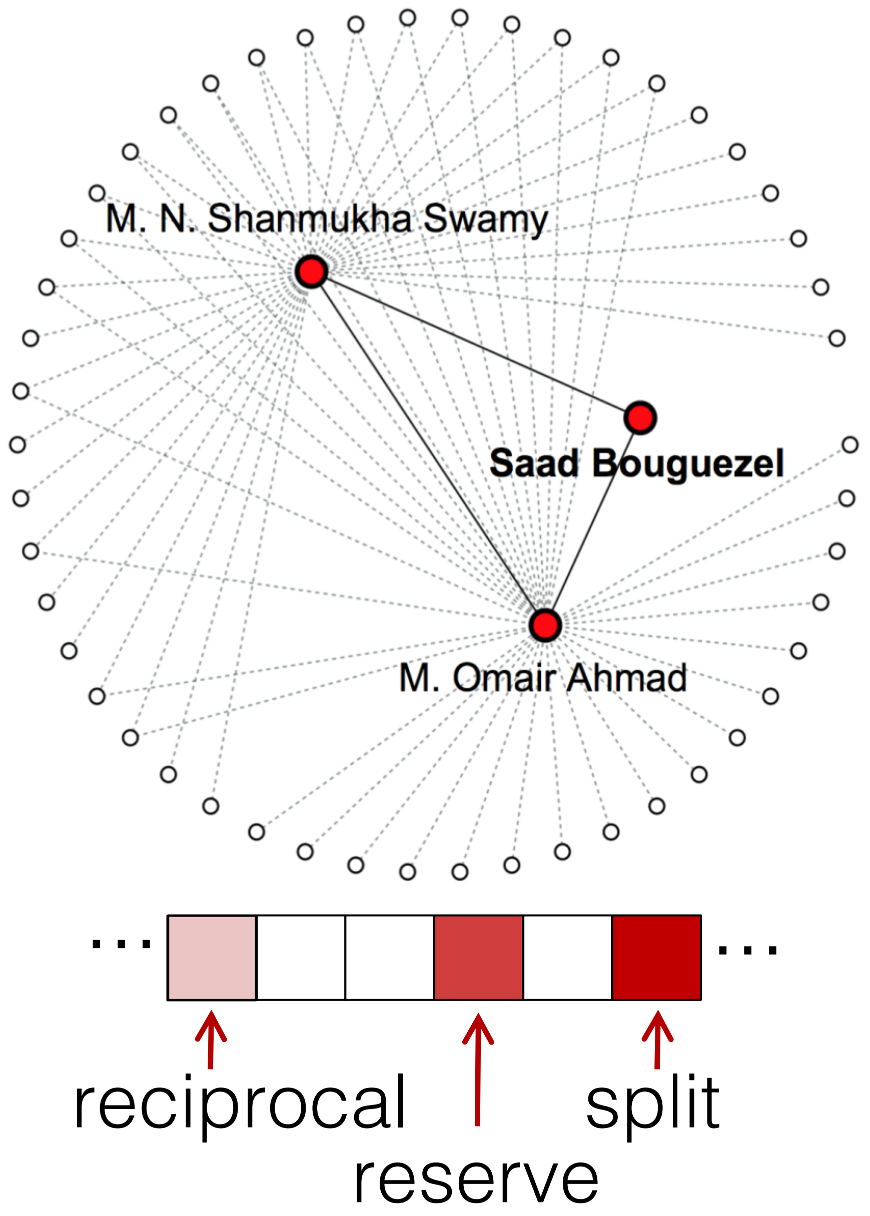
3

Anomalous groups (collective): AMEN

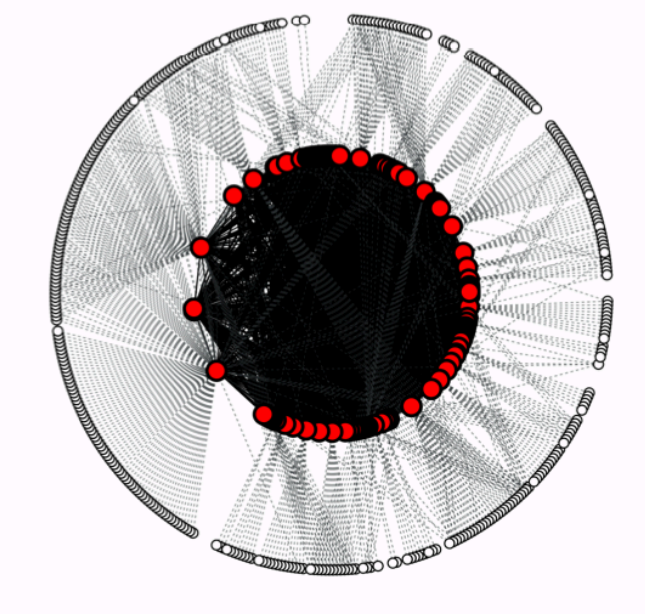
telescopic op-amps



split-radix FFT

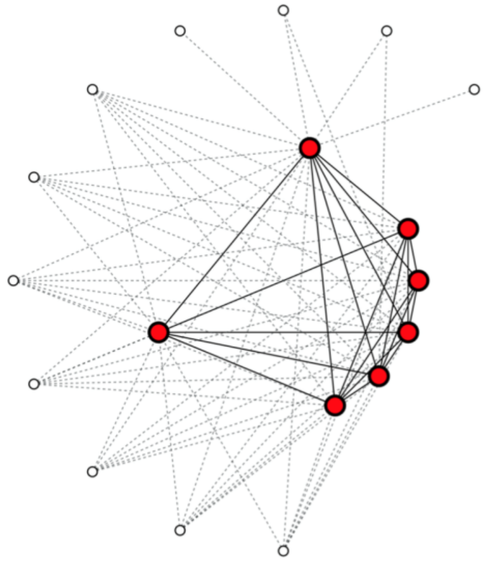


Anomalous groups (collective): AMEN



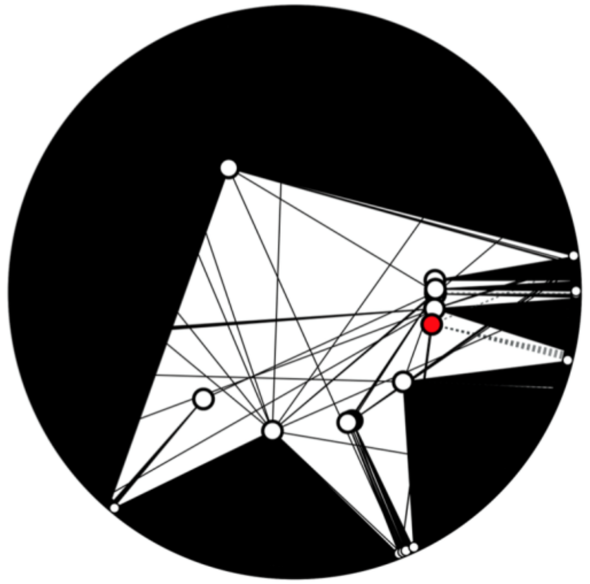
DBLP

$L_1 = 0.979,\ L_2 = 2.17$



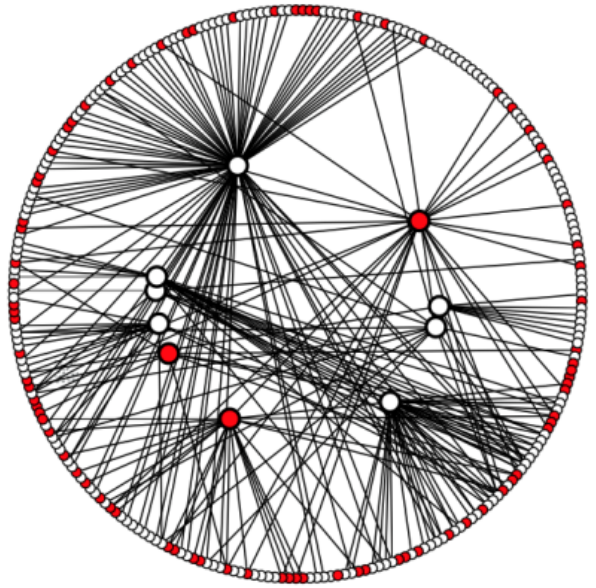
Twitter

$L_1 = 0.724,\ L_2 = 1.10$



Citeseer

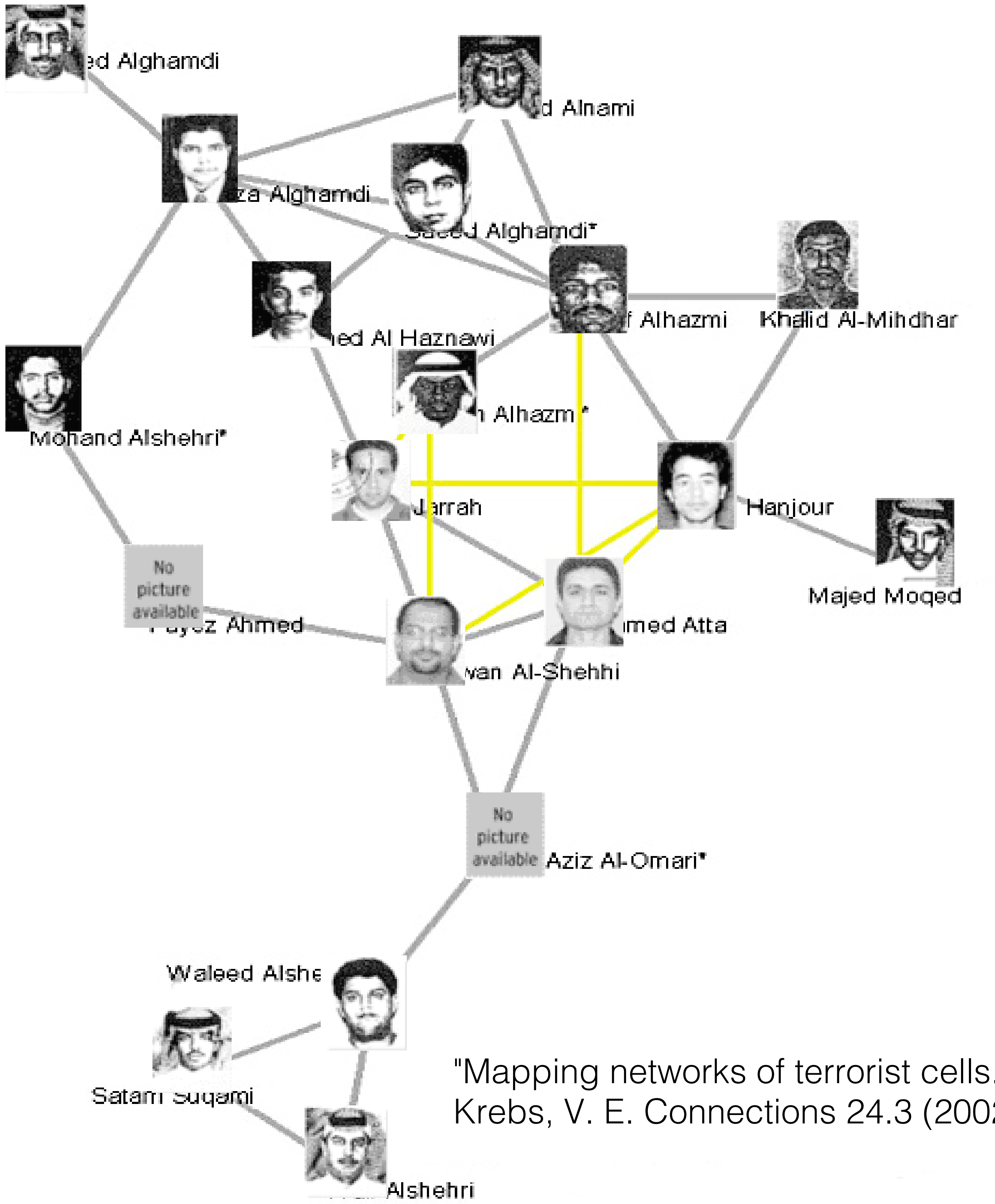
$L_1 = L_2 = -0.956$



Google+

$L_1 = L_2 = -0.873$

Anomalous groups in malice detection

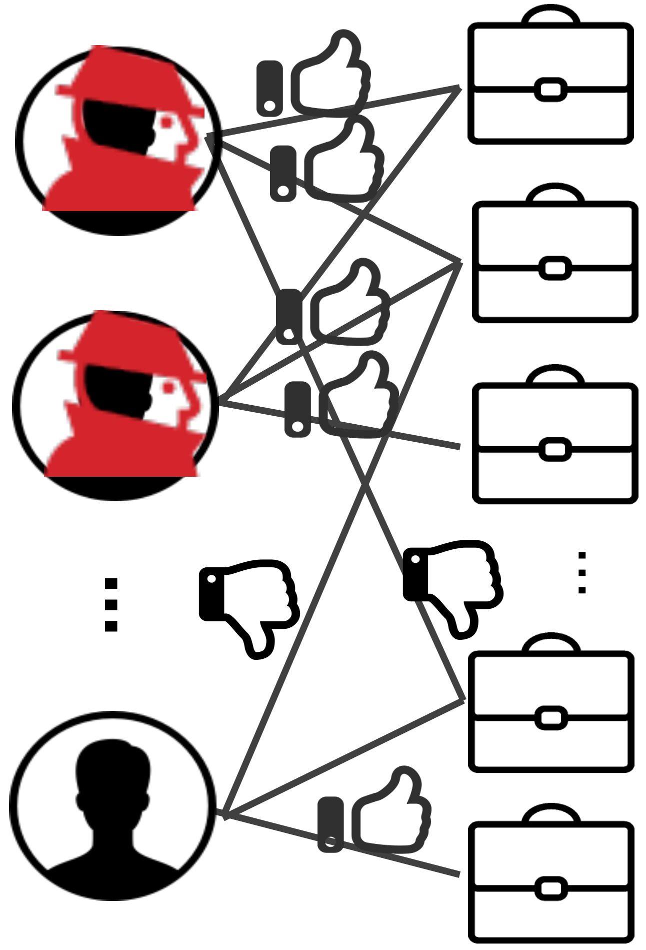


"Mapping networks of terrorist cells."
Krebs, V. E. Connections 24.3 (2002): 43-52

- In other contexts, "too dense"ly connected groups may be indicative of malice/fraud
- 9/11 hijackers were **densely linked** via
  - kinship
  - school/training
  - travel/financial records
  - meetings
  - …

Anomalous groups in malice detection

- **Opinion fraud**: Groups of users promoting/demoting businesses

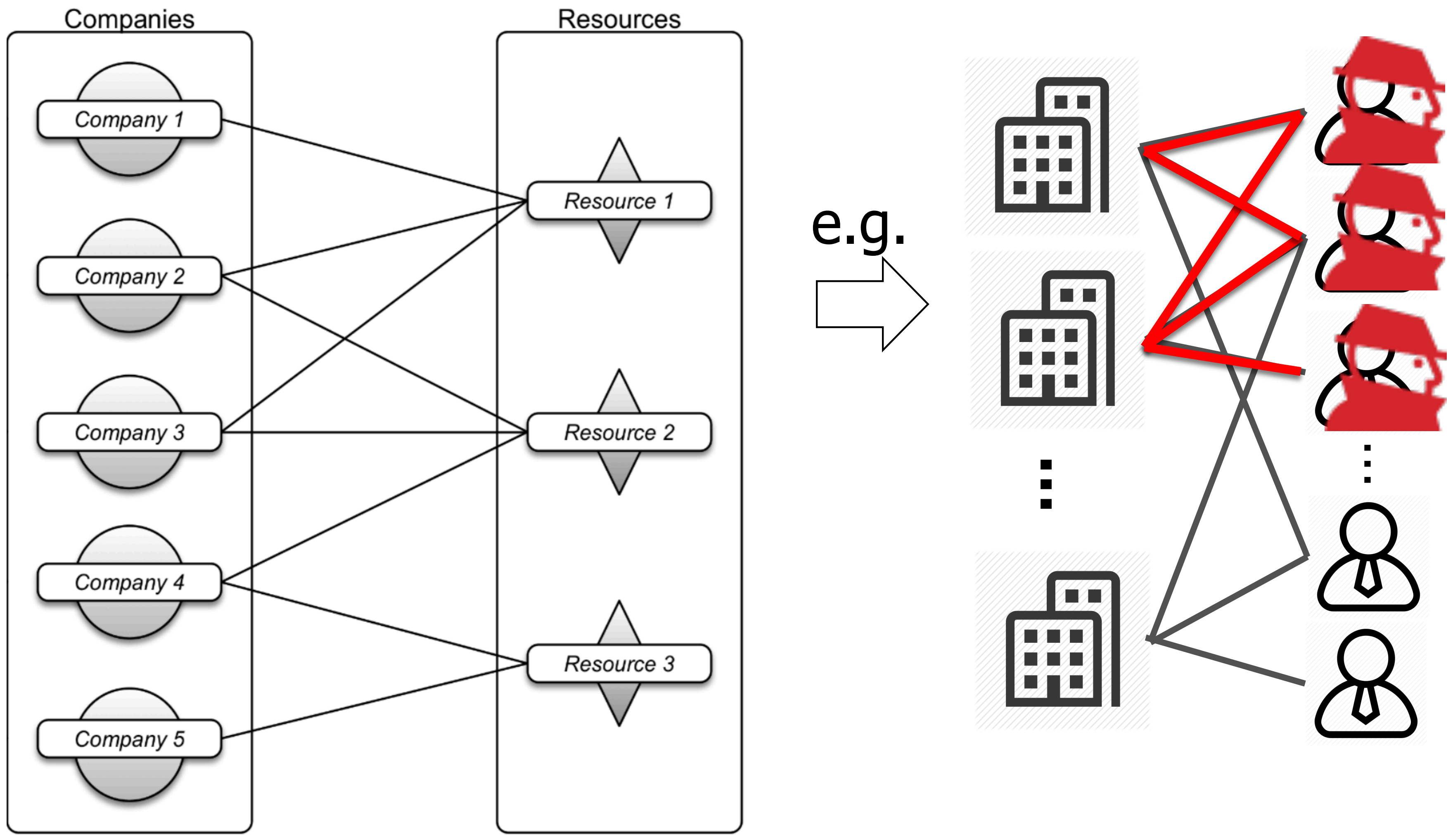


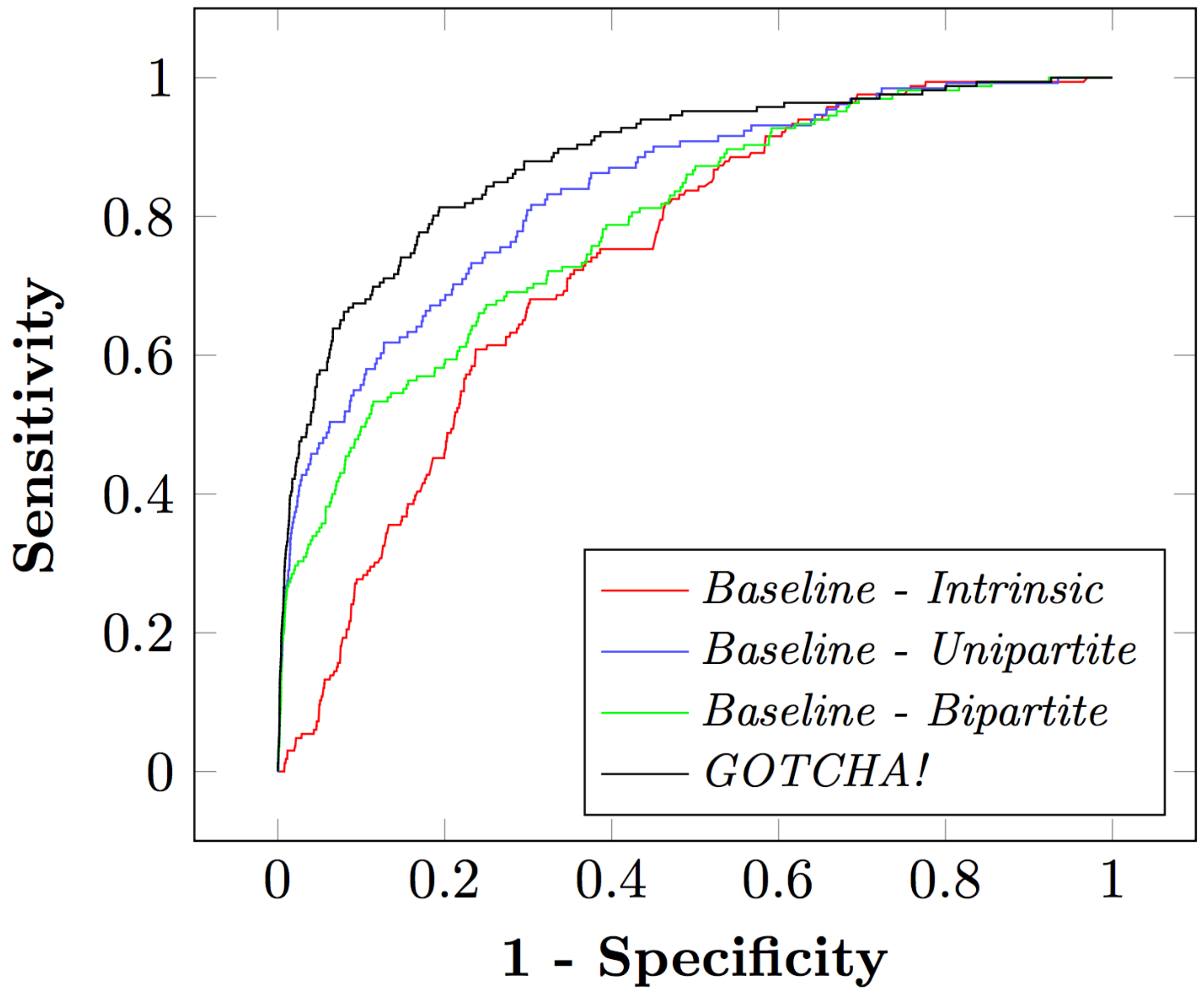
	Review Ranking					
	AP			AUC		
	Y'Chi	Y'NYC	Y'Zip	Y'Chi	Y'NYC	Y'Zip
RANDOM	0.1327	0.1028	0.1321	0.5000	0.5000	0.5000
FRAUDEAGLE	0.1067	0.1122	0.1524	0.3735	0.5063	0.5326
WANG ET AL.	0.1518	0.1255	0.1803	0.5062	0.5415	0.5982
PRIOR	0.2241	0.1789	0.2352	0.6707	0.6705	0.6838
SPEAGLE	**0.3236**	**0.2460**	**0.3319**	**0.7887**	**0.7695**	**0.7942**

Opinion Fraud Detection in Online Reviews using Network Effects. [Akoglu+] ICWSM, 2013
Collective Opinion Spam Detection: Bridging Review Networks and Metadata. [Rayana & Akoglu] ACM SIGKDD 2015
Discovering Opinion Spammer Groups by Network Footprints. [Ye & Akoglu] ECML PKDD 2015
BIRDNEST: Bayesian Inference for Ratings-Fraud Detection. [Hooi+] SIAM SDM 2016
Collective Opinion Spam Detection using Active Inference. [Rayana & Akoglu] SIAM SDM 2016

Anomalous groups in malice detection

- **Social securities tax fraud**: Groups of resources transferred between "shadow" companies





GOTCHA! Network-based Fraud Detection for Social Security Fraud. [Van Vlasselaer+] Management Science, 63 (9), 2016

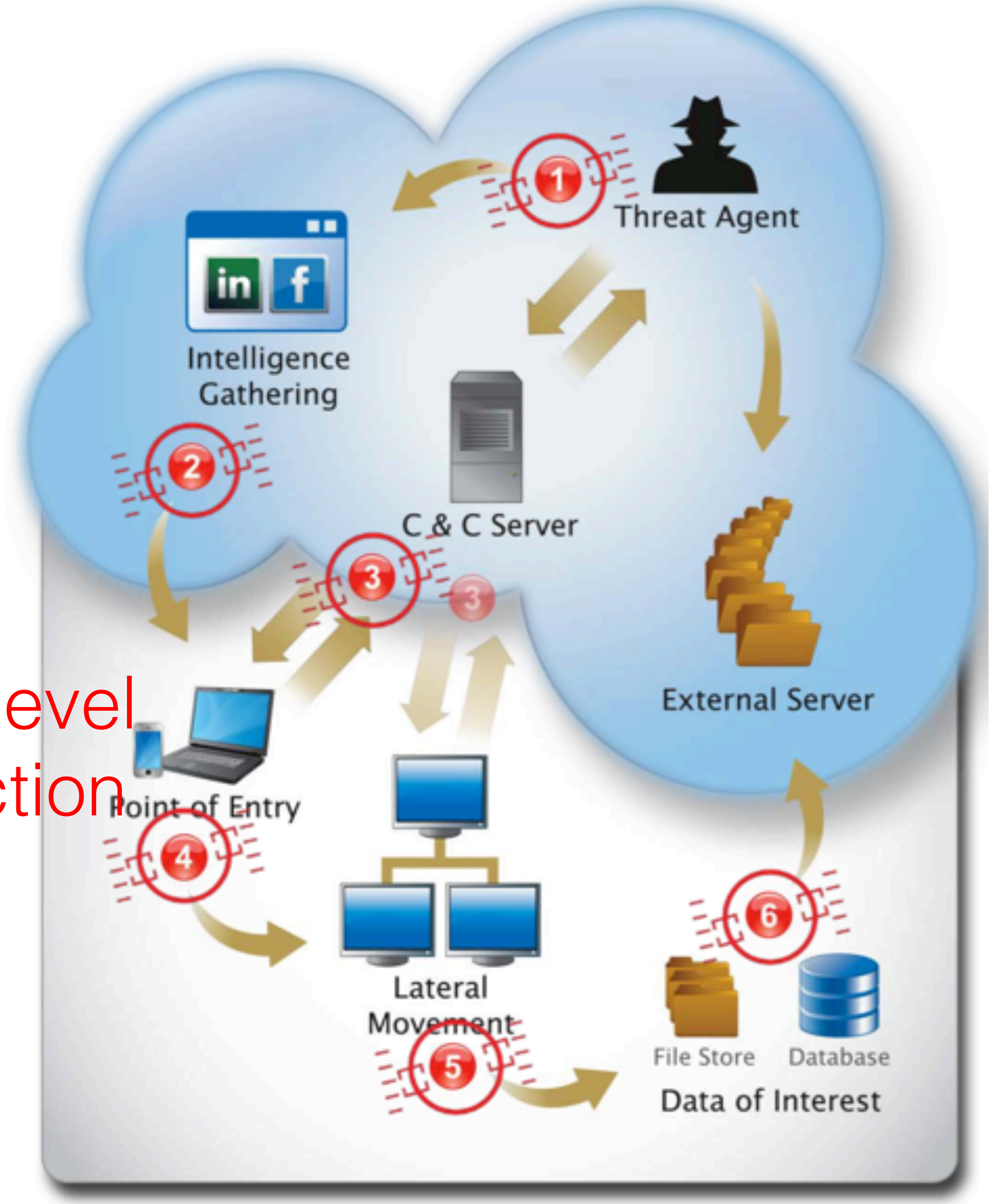
Outline

- Anomaly Detection: Motivation, Formalism, Challenges
- Graph-based Anomaly Detection
  - General-purpose (single graph)
    - Global – anomalous nodes
    - Local – group anomalies
    - Collective – anomalous groups

  - Specialized (graph database)
- Recent Trend: Deep Anomaly Detection

Anomalous graphs (System security)

- Advanced Persistent Threat



Host-level detection

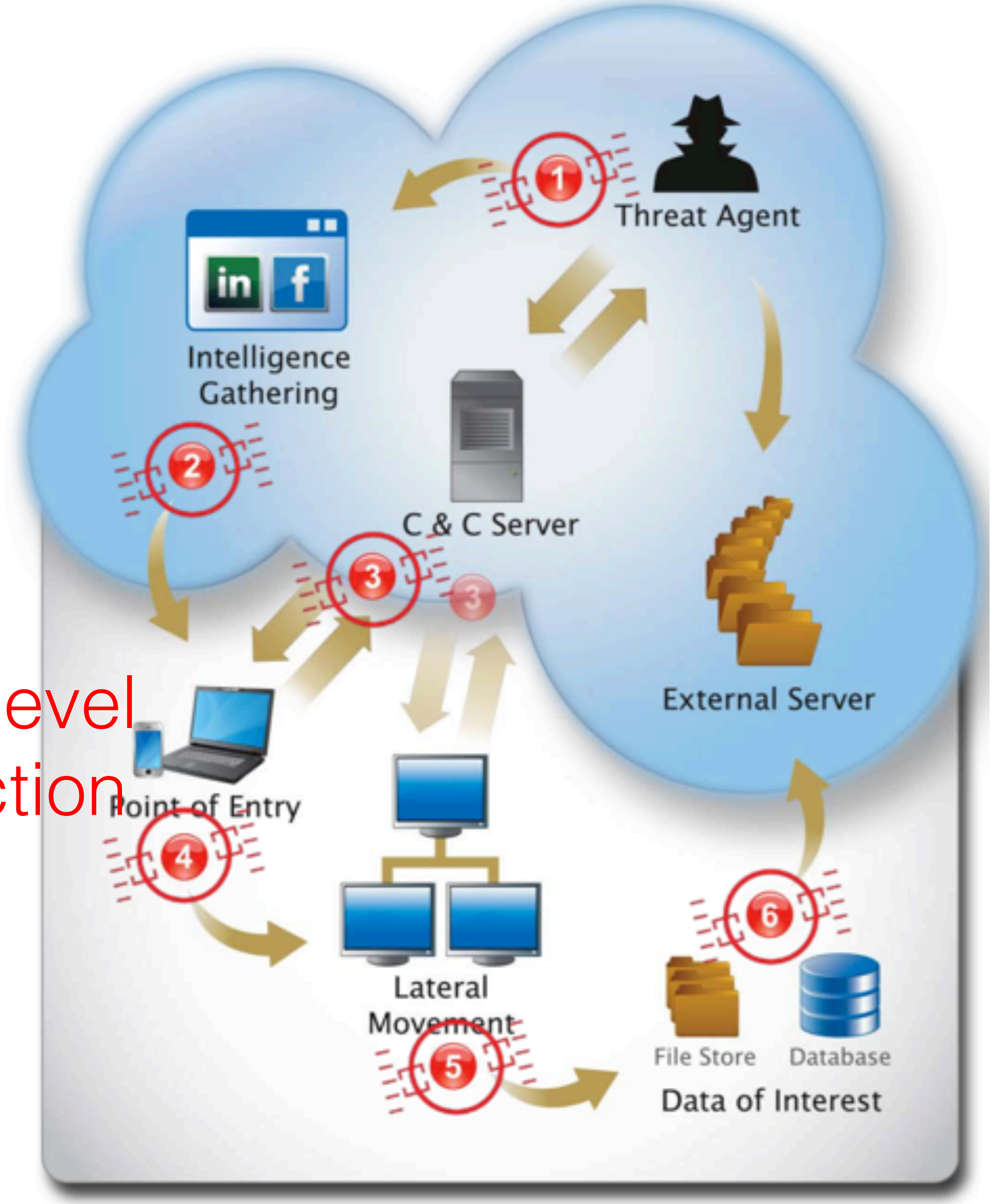
Problem Setting :

- **Given** a stream of event logs
- **Find** anomalous system events

time	pid	event	arg/data
100	10639	fork	NULL
200	10640	execve	/bin/sh
300	10650	read	STDIN
400	10640	fstat	0xbfc5598
500	10660	sock_wr	0.0.0.0
…	…	…	…

Anomalous graphs (System security)

- Advanced Persistent Threat



Host-level detection

Problem Setting :

- **Given** a stream of event logs
- **Find** anomalous system events

Given <DATA>, Find <ANOMALIES>
s.t. <CONSTRAINTS>

Requirements :

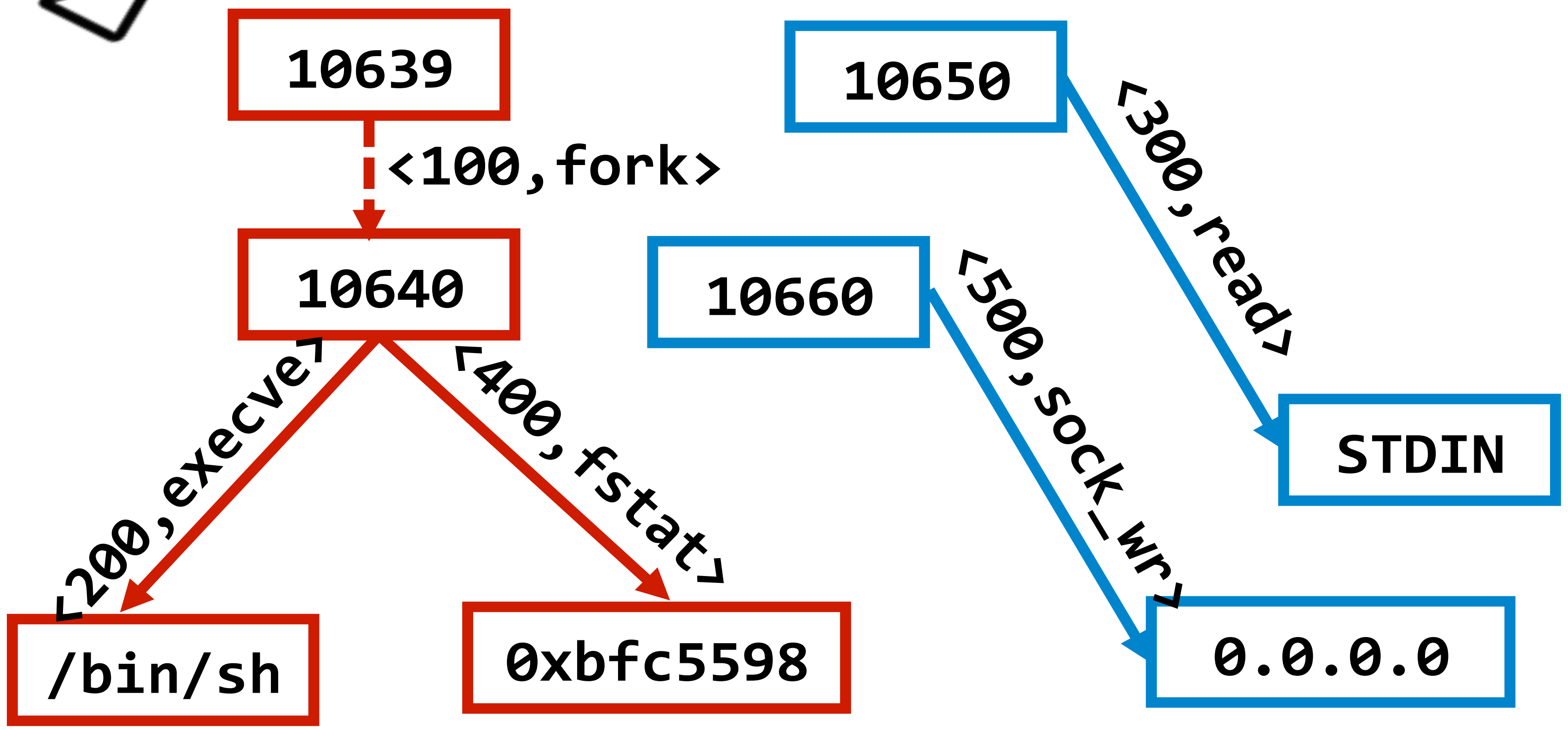
- Real-time detection
- Low-latency
- Low computational overhead
- Low memory usage

Anomalous graphs (System security)

- Each event associated with a logical flow (**tag**)

time	pid	event	arg/data	tag
100	10639	fork	NULL	1
200	10640	execve	/bin/sh	1
300	10650	read	STDIN	2
400	10640	fstat	0xbfc5598	1
500	10660	sock_wr	0.0.0.0	2
…	…	…	…	…

- Each event as a directed edge :



- Events from different flows may interleave

! Many, simultaneously-growing node&edge-labeled graphs

! Universe of labels unknown

Fast Memory-Efficient Anomaly Detection in Streaming Heterogeneous Graphs. [Manzoor+] ACM SIGKDD, 2016.

Anomalous graphs (Accounting)

- Double-entry Bookkeeping

example journal entry:

GL_Account_Number	CA_FS_Caption	Cr/Db	GL_Reporting_Amount
40020000 (Revenue)	Gross Sales (GSL)	C	-7250
40020001 (Revenue)	Gross Sales (GSL)	C	-2500
20830000 (Liabilities)	Sales Tax Payables (STP)	C	-794.63
10390000 (Assets)	Accounts Receivable (ARV)	D	10544.63

Problem Setting :

- **Given** millions of journal entries
- **Find** anomalies

(entry errors, misconduct, etc.)

Given <DATA>, Find <ANOMALIES>
s.t. <CONSTRAINTS>

L. Akoglu

Anomalous graphs (Accounting)

- Double-entry Bookkeeping example journal entry:

GL_Account_Number	CA_FS_Caption	Cr/Db	GL_Reporting_Amount
40020000 (Revenue)	Gross Sales (GSL)	C	-7250
40020001 (Revenue)	Gross Sales (GSL)	C	-2500
20830000 (Liabilities)	Sales Tax Payables (STP)	C	-794.63
10390000 (Assets)	Accounts Receivable (ARV)	D	10544.63

Problem Setting :

- **Given** millions of journal entries
- **Find** anomalies
  (entry errors, misconduct, etc.)

Given <DATA>, Find <ANOMALIES>
s.t. <CONSTRAINTS>

Requirements :

- Explainability (audit)

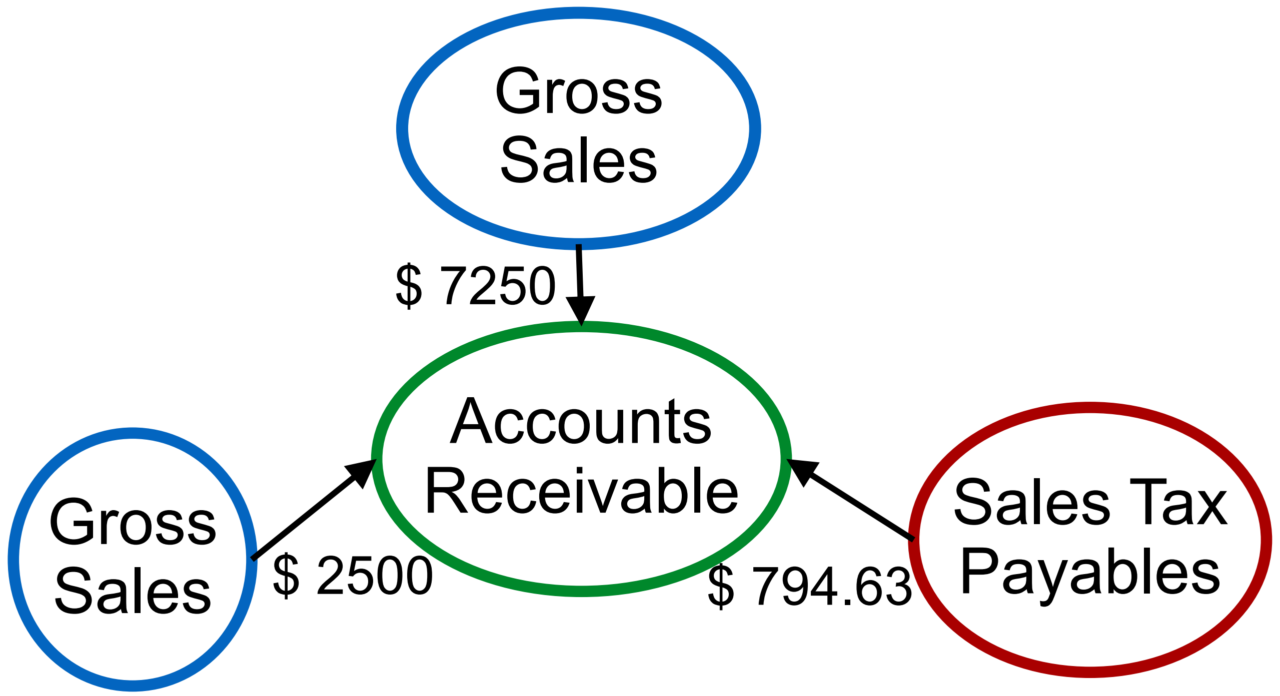
Anomaly Detection in Large Labeled Multi-Graph Databases. [Nguyen+] arXiv:2010.03600, 2020.

Anomalous graphs (Accounting)

- Transaction graphs:

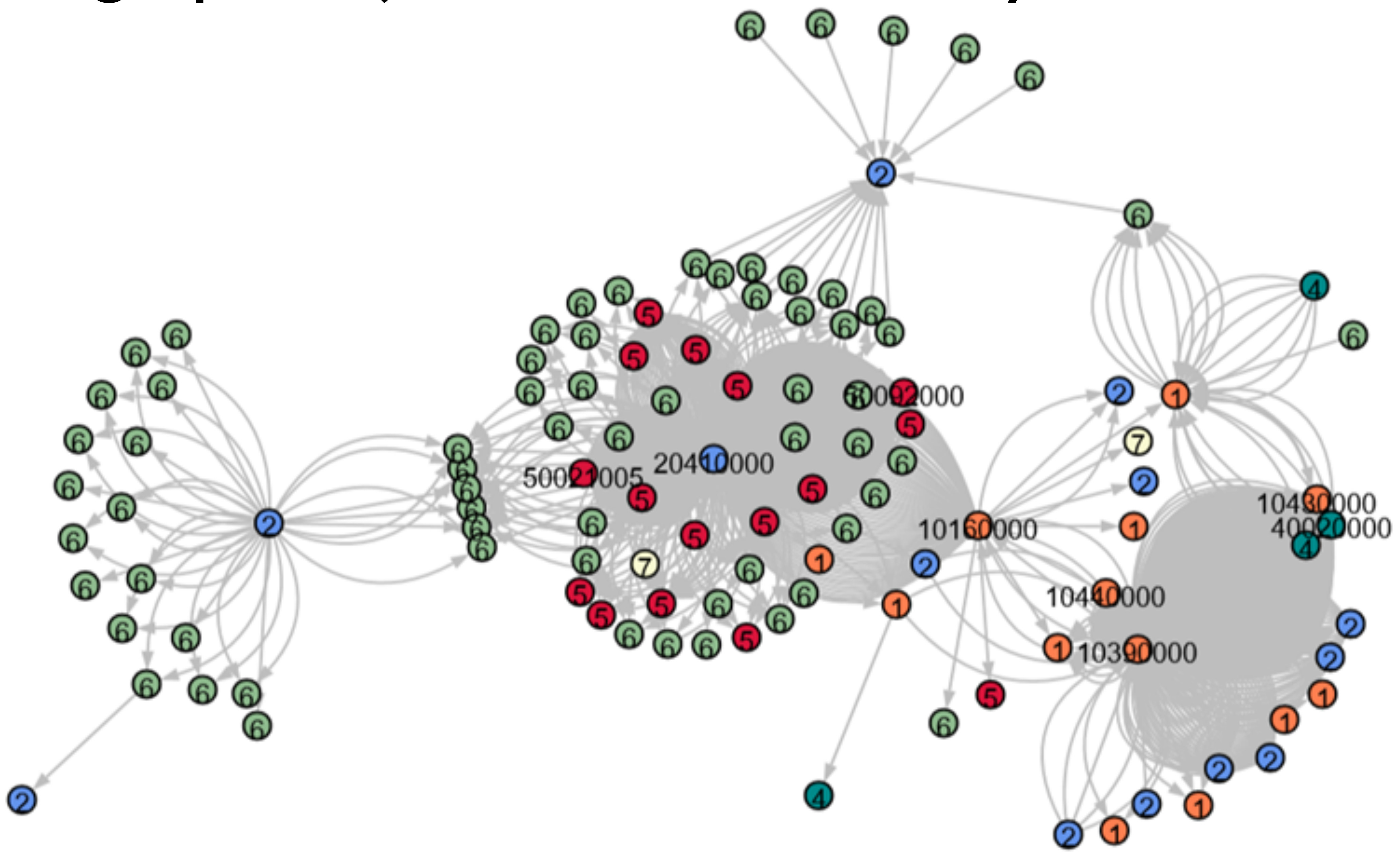
GL_Account_Number	CA_FS_Caption	Cr/Db	GL_Reporting_Amount
40060000 (Revenue)	Gross Sales (GSL)	C	-1575.00
10415000 (Assets)	Accounts Receivable (ARV)	D	1575.00

GL_Account_Number	CA_FS_Caption	Cr/Db	GL_Reporting_Amount
40020000 (Revenue)	Gross Sales (GSL)	C	-7250
40020001 (Revenue)	Gross Sales (GSL)	C	-2500
20830000 (Liabilities)	Sales Tax Payables (STP)	C	-794.63
10390000 (Assets)	Accounts Receivable (ARV)	D	10544.63



Anomalous graphs (Accounting)

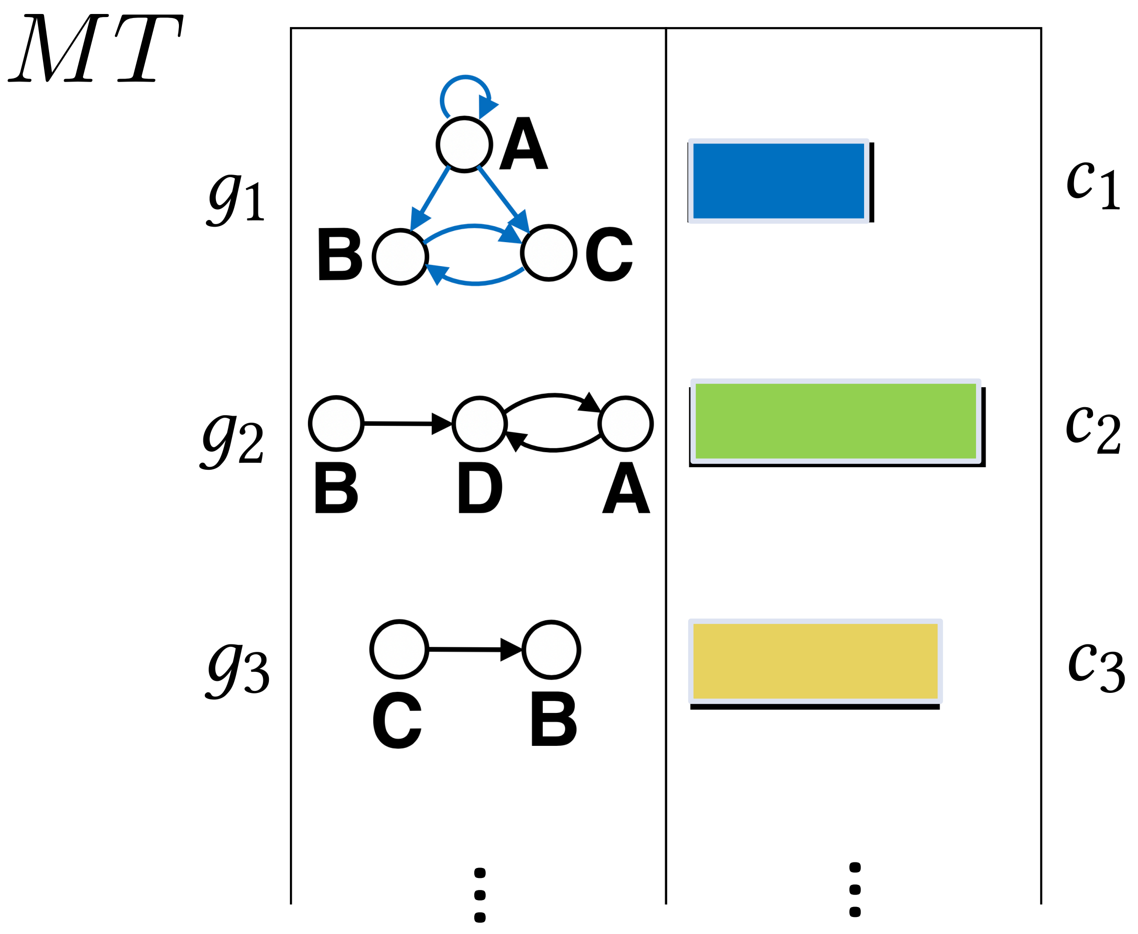
- Transaction graph of journals over 10-day window:

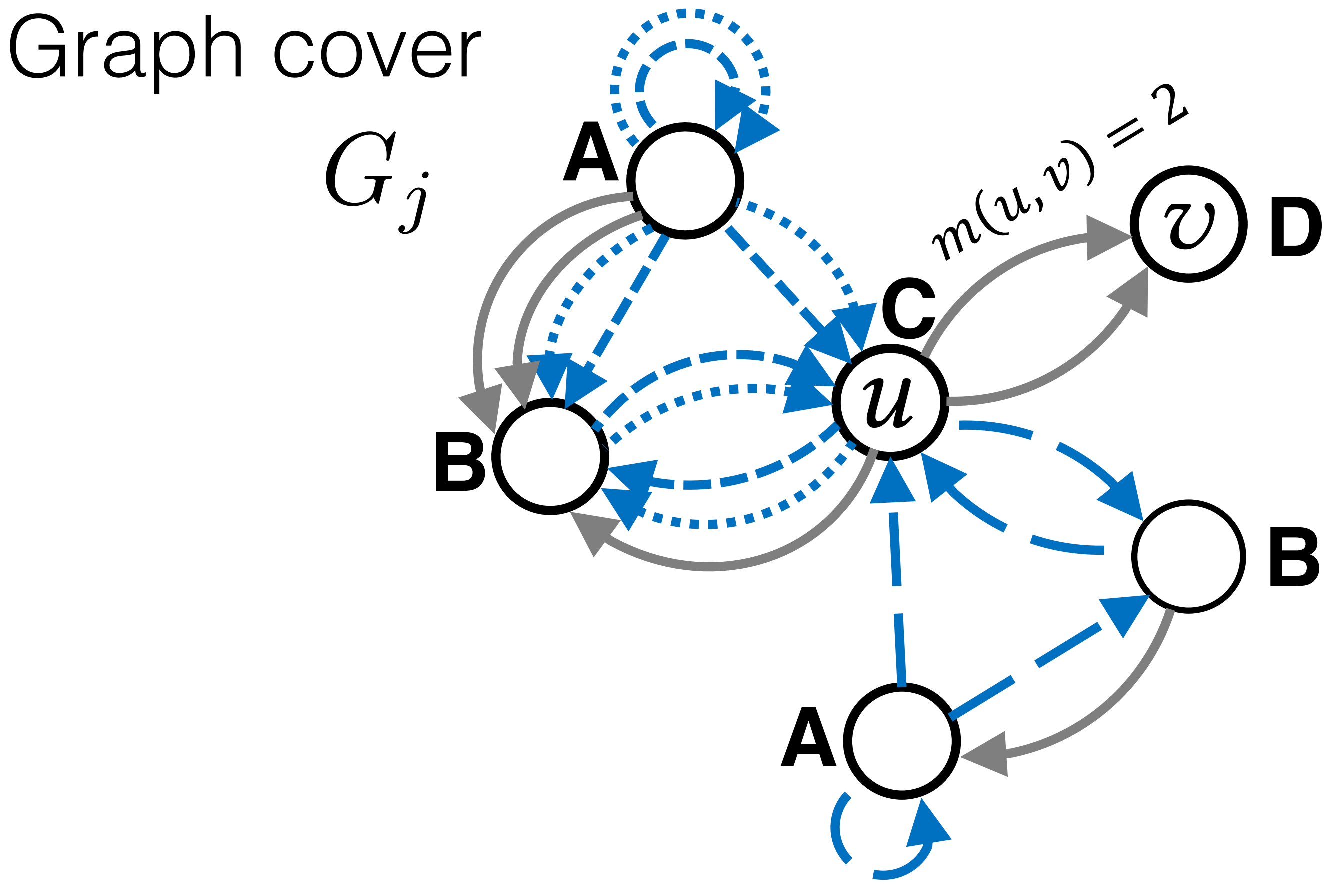


Anomalous graphs (Accounting)

- Anomaly detection via data description/encoding

$$\underset{MT \subseteq \mathcal{MT}}{\text{minimize}} \quad L(MT, \mathcal{G}) = \underbrace{L(MT)}_{\text{model code length}} + \underbrace{L(\mathcal{G}|MT)}_{\text{data code length}} \quad , \quad (1)$$



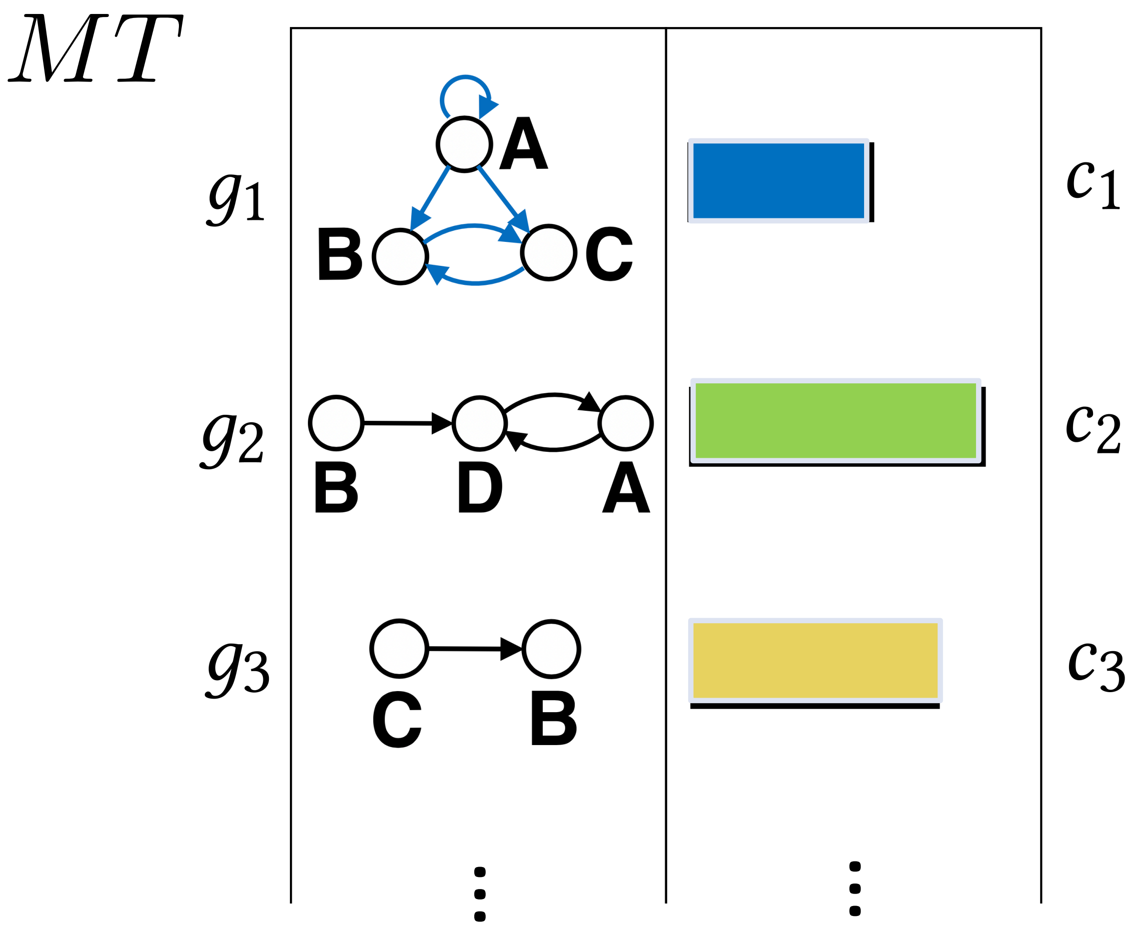


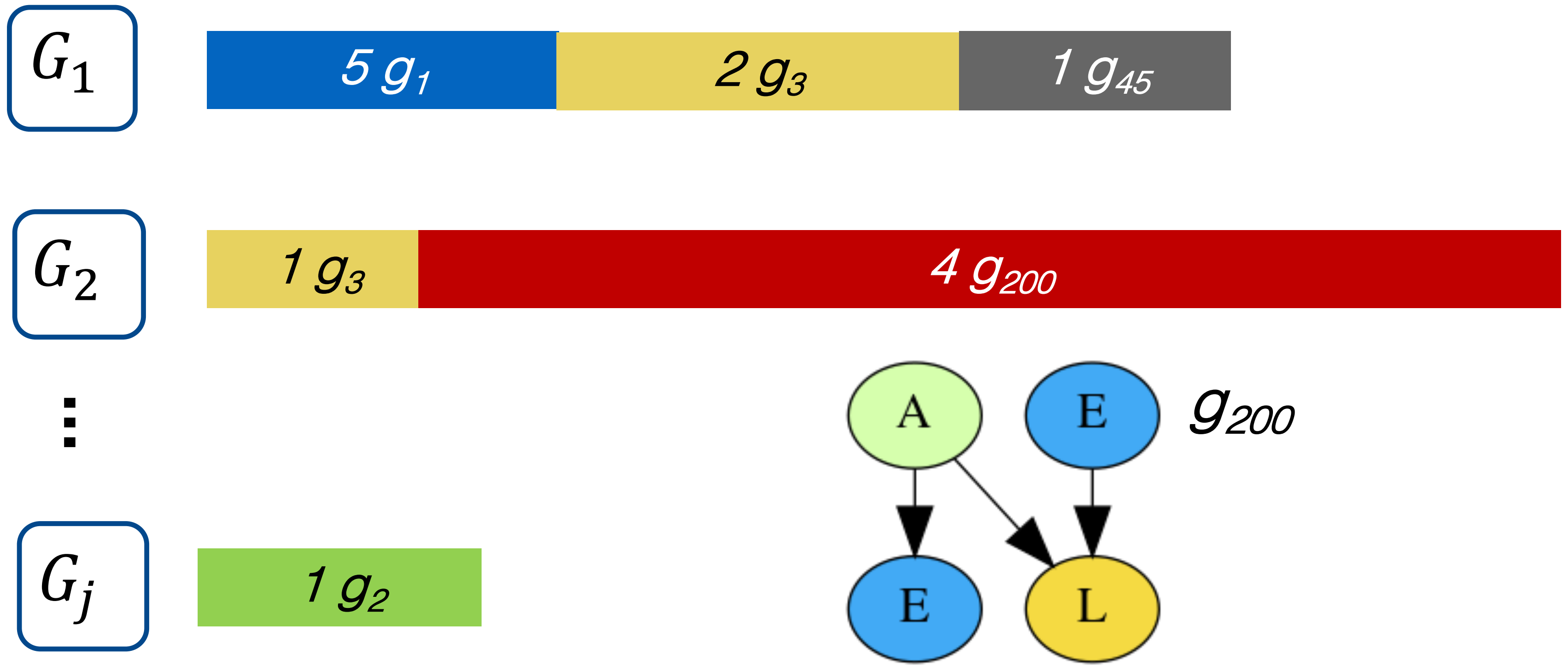
L. Akoglu
Carnegie Mellon

Anomalous graphs (Accounting)

- Anomaly detection via data description/encoding

$$\underset{MT \subseteq \mathcal{MT}}{\text{minimize}} \quad L(MT, \mathcal{G}) = \underbrace{L(MT)}_{\text{model code length}} + \underbrace{L(\mathcal{G}|MT)}_{\text{data code length}} \,, \quad (1)$$





Anomalous graphs (Accounting)

- Anomaly detection via data description/encoding

Path injections

Method	Prec@10	Prec@100	Prec@1000	AUC	AP
CODETECT	**0.900**	**0.990**	**0.999**	**0.995**	**0.772**
SMT	0.600	0.440	0.784	0.906	<u>0.733</u>
SUBDUE	<u>0.800</u>	0.710	0.685	0.930	0.555
GF+IFOREST	0.400	0.230	0.497	0.959	0.429
G2V+IFOREST	0.000	0.100	0.819	0.824	0.380
DGK+IFOREST	0.300	0.140	0.023	0.858	0.097

Type perturbations

Method	Prec@10	Prec@100	Prec@1000	AUC	AP
CODETECT	**0.800**	**0.720**	**0.709**	0.918	0.359
SMT	0.000	0.100	0.174	0.883	0.192
SUBDUE	<u>**0.800**</u>	<u>0.710</u>	<u>0.685</u>	**0.920**	**0.555**
GF+IFOREST	0.200	0.080	0.027	0.832	0.092
G2V+IFOREST	0.000	0.030	0.030	0.499	0.030
DGK+IFOREST	0.100	0.030	0.038	0.801	0.074

Outline

- Anomaly Detection: Motivation, Formalism, Challenges
- **Graph-based Anomaly Detection**
  - General-purpose (single graph)
    - Global – anomalous nodes
    - Local – group anomaly
    - Collective – anomalous groups
  - Graph-level anomalies

Recent Trend: Deep Anomaly Detection

Deep Anomaly Detection

- Representation learning: transformative for applications in NLP/translation, recommender systems, etc.

- Why not **automatically learn data representations for anomaly detection**?

  Deep Learning for Anomaly Detection: A Survey. [Chalapathy & Chawla] Jan. 2019

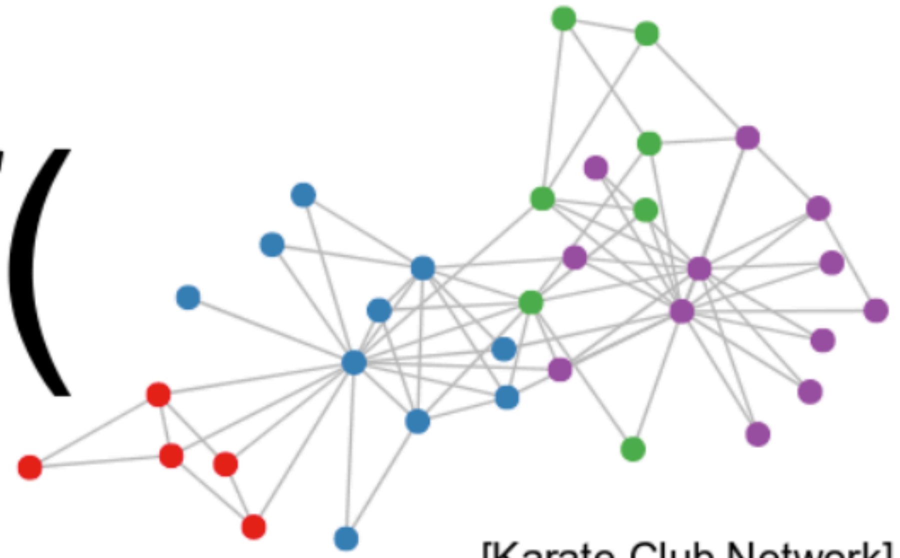
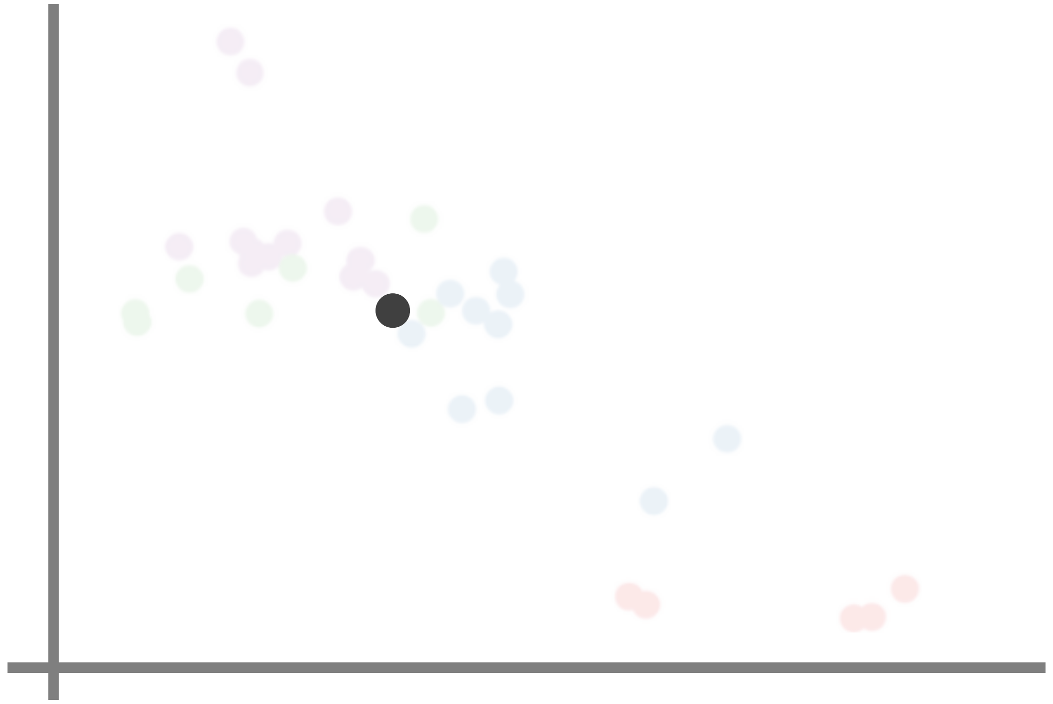
  Deep Learning for Anomaly Detection: A Review. [Pang+] July 2020

  A Unifying Review of Deep and Shallow Anomaly Detection. [Ruff+] Sep. 2020

- Ideas easily transfer to graph data

Deep Anomaly Detection

Graph Embedding

f() = 

[Karate Club Network]

Outlier Detector

Deep Anomaly Detection

Graph Embedding

☺ Can seamlessly **handle various types of graphs**: labeled, attributed, multi-edges, weights

☺ Can do **end-to-end** learning (one-class, reconstruction)

☹ Embeddings capture **general prevalent patterns**, may not be suitable for anomaly detection

☹ **Hyper-parameter tuning** becomes key for success!

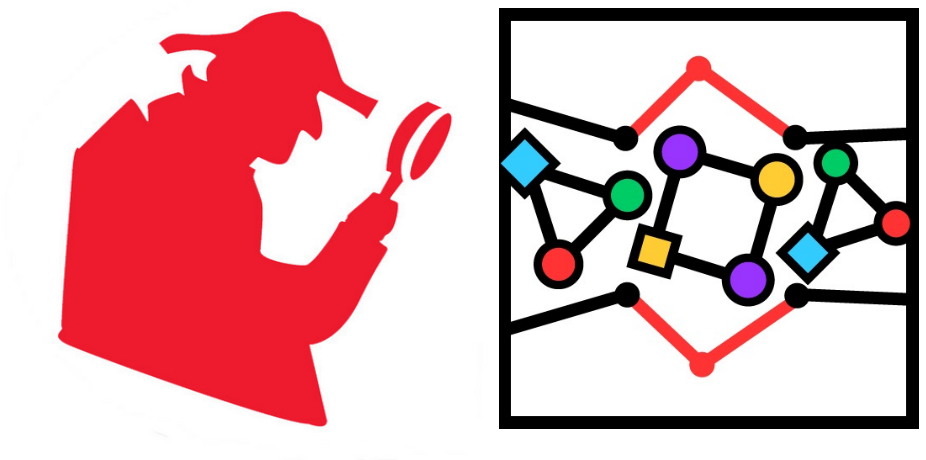
Unsupervised model selection – likely a critical future direction

Graph-based Anomaly Detection

Code, Data, Papers, Slides

www.cs.cmu.edu/~lakoglu/

http://www.andrew.cmu.edu/user/lakoglu/pubs.html#code



Thanks!