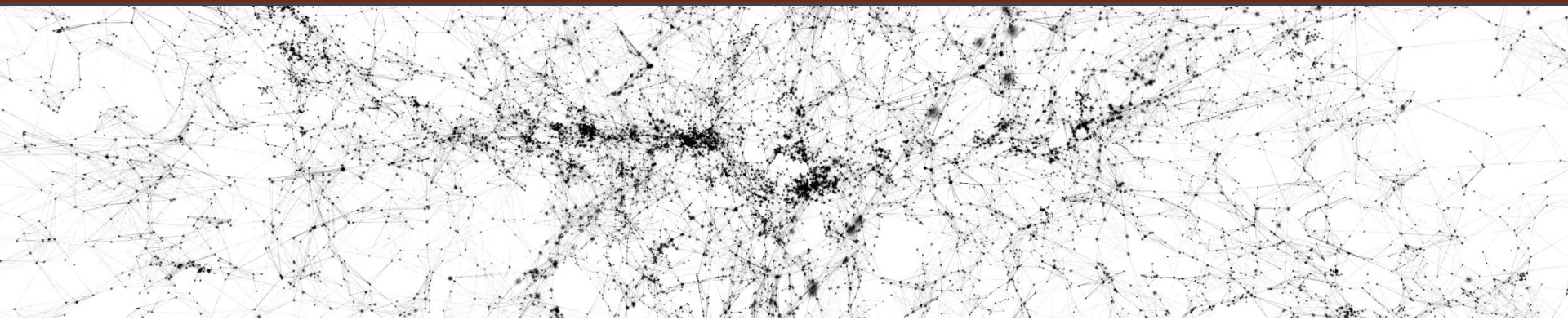
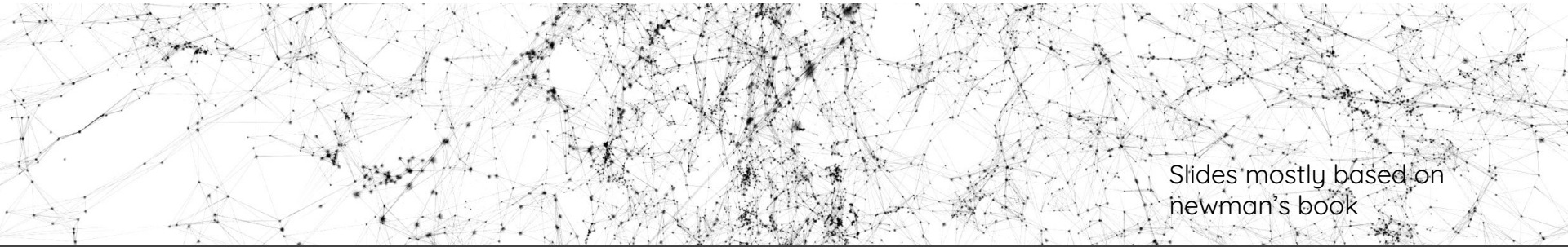


Anomalies

Analysis of complex interconnected data



Slides mostly based on newman's book

Outline

- Quick Notes
- Quick Recap of Modules
- Anomalies
  - Anomaly detection challenges and applications
  - Feature based anomaly detection (not-normal)
    - Oddball example
    - SOAR example in attributed graphs
    - Lookout example in explaining anomalies
  - Dense block detection (coordinated)
  - Other approaches

Quick Notes

- Reminder second assignment is out, due on Oct 4th
  - http://www.reirab.com/Teaching/NS20/Assignment_2.pdf
  - Submit 2 files (report.pdf, code.zip) as a Group (pairs or two or individual) in Mycourses
- We start seminars from Oct 7th. Please sign up for the desired paper, slots
  - The sooner you put down your name, more options you have to choose from
  - Please put your name for **two** slots, do not change the entries that are already booked
  - Find the link to editable spreadsheet in slack or Mycourses
- Any questions?

Deadlines

- assignment 1 due on Sep. 20th
- assignment 2 due on Oct. 4th
- assignment 3 due on Oct. 18th
- project proposal slides due on Oct. 25th
- project proposal due on Nov. 1th
- Reviews (first round) due on Nov. 8th
- project progress report due on Nov. 22nd
- Reviews (second round) due on Nov. 29th
- project final report slides due on Dec. 1st
- project final report due on Dec. 6th
- Reviews (third round) due on Dec. 13th
- project revised report and rebuttal due on Dec. 20th
- note: dates are tentative, please check them for the updated deadlines

Quick Recap of Modules

- Real graphs are modular and finding modules (a.k.a communities) has myriad of applications
- Spectral clustering finds the sparsest balanced cut, using the signs on elements in the eigenvector associated to the second smallest eigenvalue of the Laplacian matrix
- Generally we can define a local (one module) or global (all modules) measure
  - Normalized cut and conductance are local, Q-modularity is global
  - Many other ways to define an objective or algorithm without an explicit objective
    - TopLeaders: K-means for graph clustering
    - Hierarchical division by removing edges with high betweenness centrality
- Modularity Optimization
  - FastModularity & Louvain, resolution limits of Modularity
- Overlap extensions: e.g.link clustering
- Evaluating clustering results
  - Validation on benchmarks with known ground-truth
  - Clustering agreement measures the dispersion in the confusion matrix

Outline

- Quick Notes
- Quick Recap of Modules
- **Anomalies**
  - Anomaly detection challenges and applications
  - Feature based anomaly detection (not-normal)
    - Oddball example
    - SOAR example in attributed graphs
    - Lookout example in explaining anomalies
  - Dense block detection (coordinated)
  - Other approaches

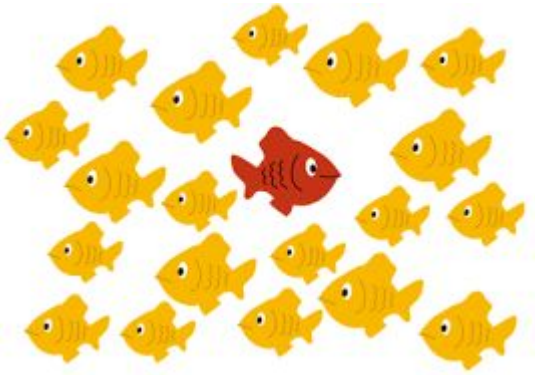
What tasks are unsupervised in Network Science?

- Community Detention
  - a.k.a. clustering nodes, finding modules
  - If semi-supervised or supervised, it becomes attribute prediction or classification
- **Anomaly Detection**
  - A.k.a. outlier detection
  - Goes hands in hands with pattern detection
- Summarization
  - How to compress the graph
- Visualization
  - How to plot the graph
- Alignment
  - How to align two given graphs

What is an anomaly?

“An outlier is an observation that differs so much from other observations as to arouse suspicion that it was generated by a different mechanism.”

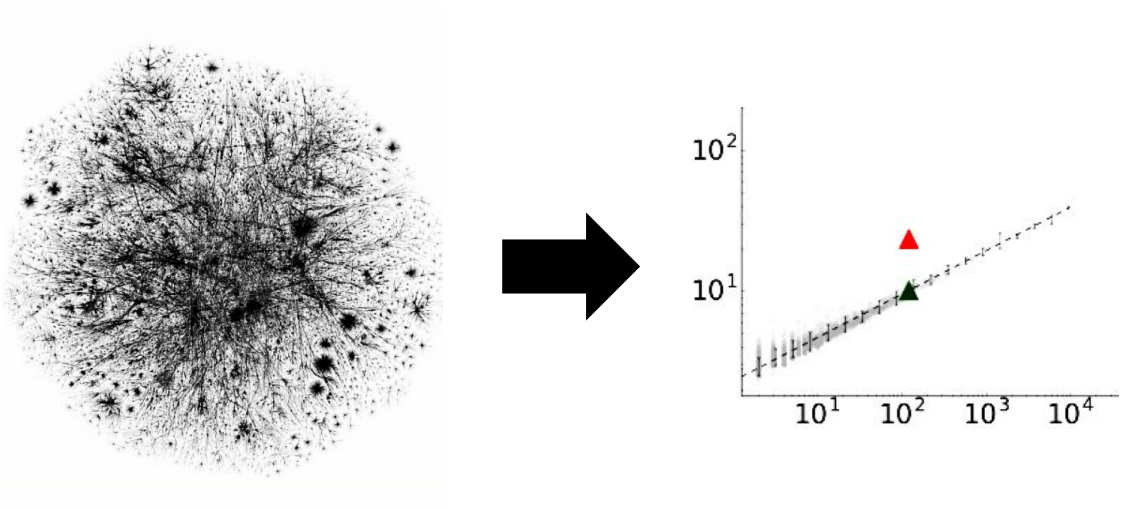
Hawkins’ Definition of Outlier, 1980



Slides based on:
Akoglu L, Tong H, Koutra D. Graph based anomaly detection and description: a survey. Data mining and knowledge discovery. 2015 May 1;29(3):626-88.

General Graph Anomaly Detection Problem

Given a graph, find the graph objects (nodes/edges/substructures) that differ significantly from the majority of the reference objects in the graph.



Challenges

Unlabeled and Hard to label ⇒ No(isy) label

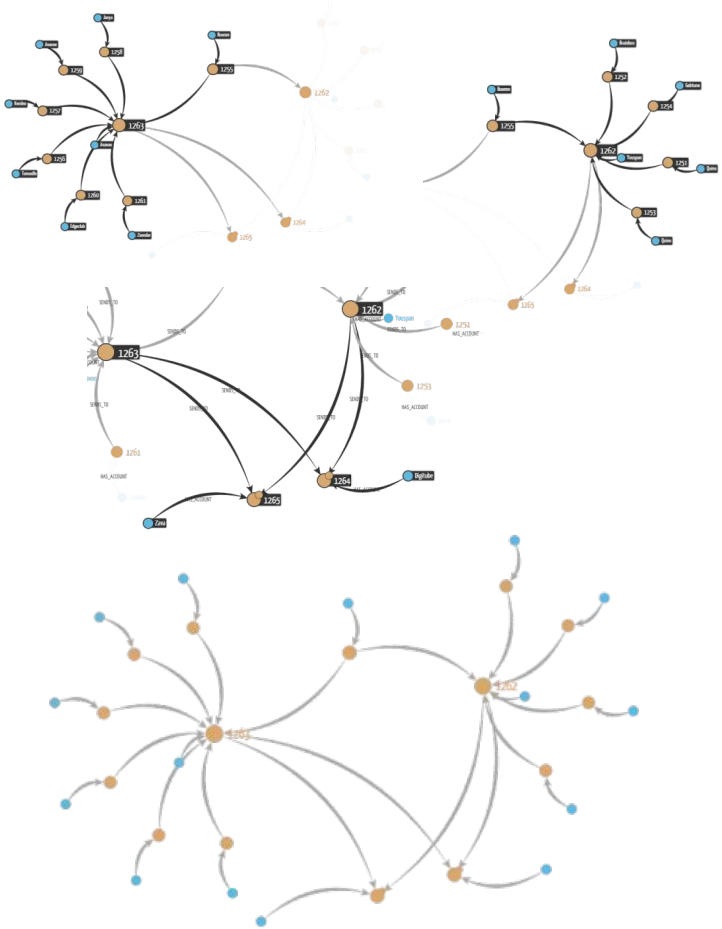
Very few anomalies ⇒ Imbalance

No fix definition ⇒ adversarial

Hard to explain

Applications

- Spam & intrusion detection
- Bot & troll detection
- Fake review, false advertising
- Fraud detection
  - Money-laundering in bank transactions
  - Phishing in blockchain transactions
  - ...

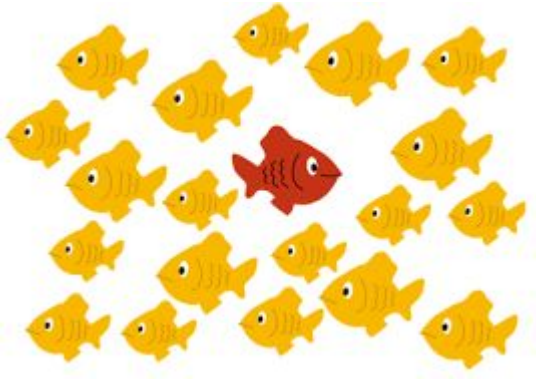


Check out this example of detecting money laundering using graphs
https://linkurio.us/blog/investigating-money-laundering-scheme/

Feature based Anomaly Detection

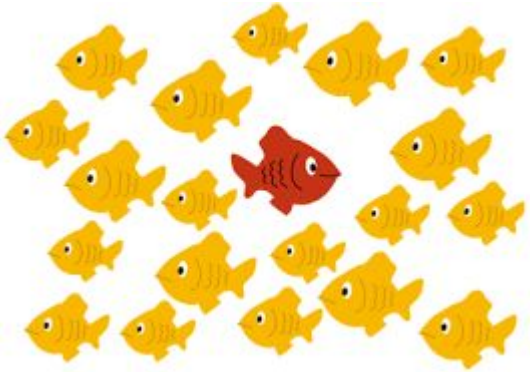
1. Extract features for each node/edge/subgraph
   features that could yield “laws”
   features fast to compute and interpret
2. Detect patterns: regularities
3. Detect anomalies: “distance” to patterns

outlier, anomaly, outbreak, event, fraud, …
rare, isolated, surprising



Feature based Anomaly Detection

1. Extract features for each node/edge/subgraph
   features that could yield “laws”
   features fast to compute and interpret
2. Detect patterns: regularities
3. Detect anomalies: “distance” to patterns



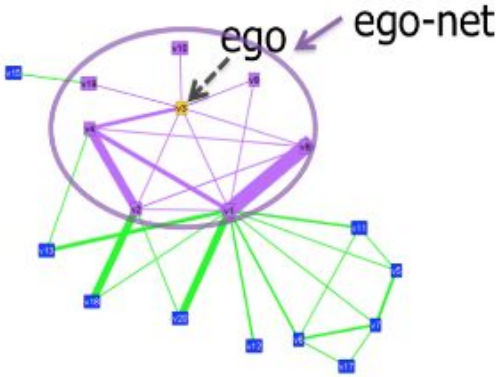
Let's look at one example based on features of **egonets**

Outline

- Quick Notes
- Quick Recap of Modules
- Anomalies
  - Anomaly detection challenges and applications
  - Feature based anomaly detection (not-normal)
    - **Oddball example**
    - SOAR example in attributed graphs
    - Lookout example in explaining anomalies
  - Dense block detection (coordinated)
  - Other approaches

Feature based Anomaly Detection Example: Oddball

1. For each node,
   a. Extract "**ego-net**" (=1-step neighborhood)
   b. Extract features (#edges, total weight, etc.)
2. Detect patterns: regularities
3. Detect anomalies: "distance" to patterns

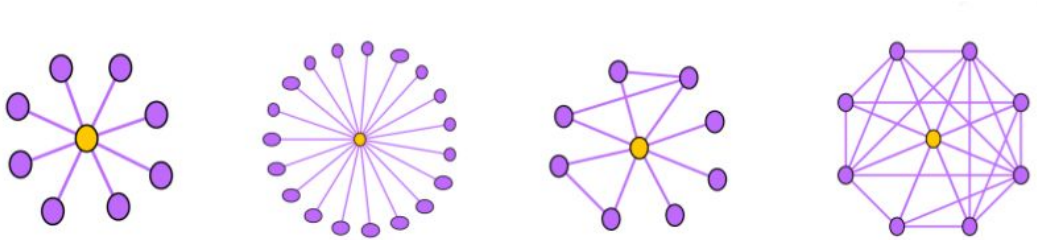


Leman Akoglu, Mary McGlohon, and Christos Faloutsos. *OddBall: Spotting anomalies in weighted graphs*. In Proceedings of the 14th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), 2010.

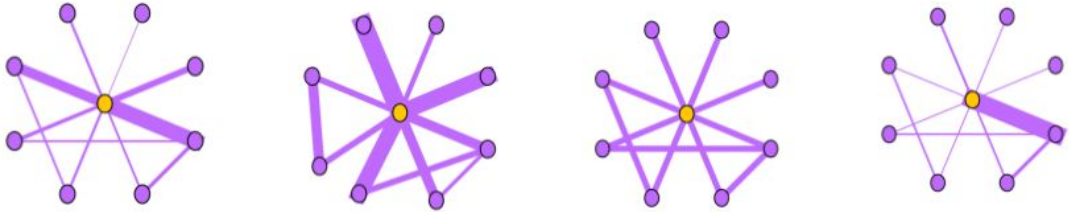
Egonet: a commonly studied substructure, 1-step (one hub) neighborhood, subgraph of all neighbours

Feature based Anomaly Detection Example: Oddball

How to categorize different egonets? What features to compute?



What if we have weights?



Slides based on L. Akoglu & C. Faloutsos, Anomaly detection in graph data (ICDM'12)

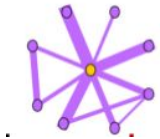
Feature based Anomaly Detection Example: Oddball

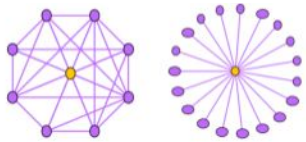
Features to characterize an egonet i:

N_i : number of nodes in egonet i (neighbors or degree of ego i)

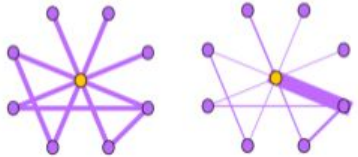
E_i : number of edges in egonet i

W_i : total weight of egonet i

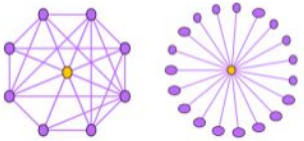




$\lambda_{w,i}$: principal eigenvalue of the weighted adjacency matrix of egonet i



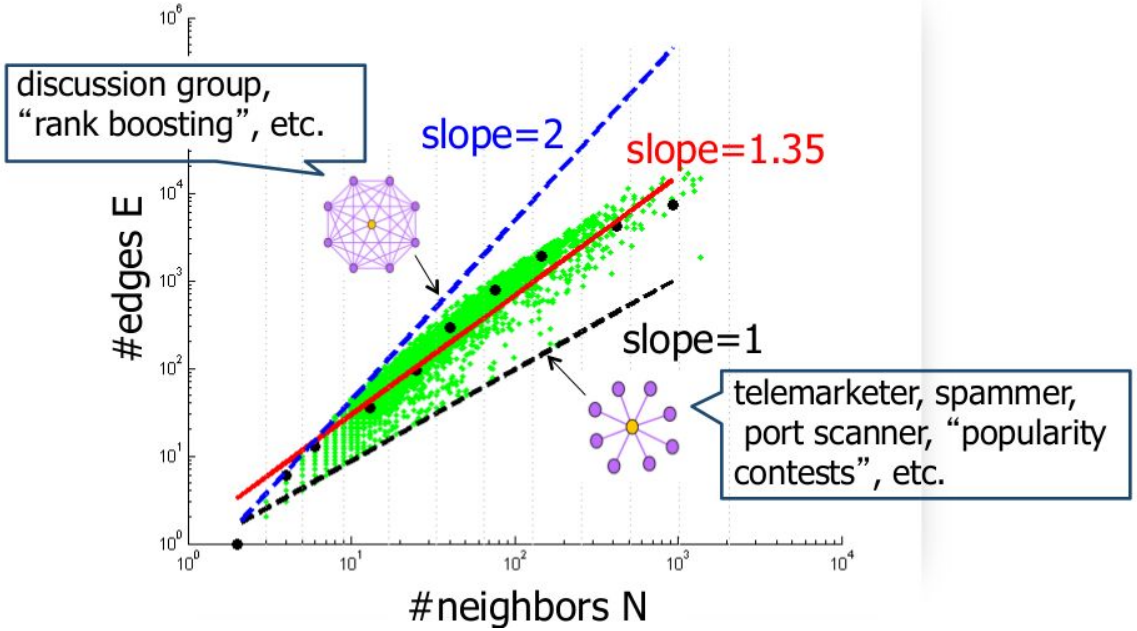
Oddball: CliqueStar Pattern



detects near-cliques and stars

Egonet Density Power Law

$E_i \propto N_i^{\alpha}, \quad 1 \le \alpha \le 2$

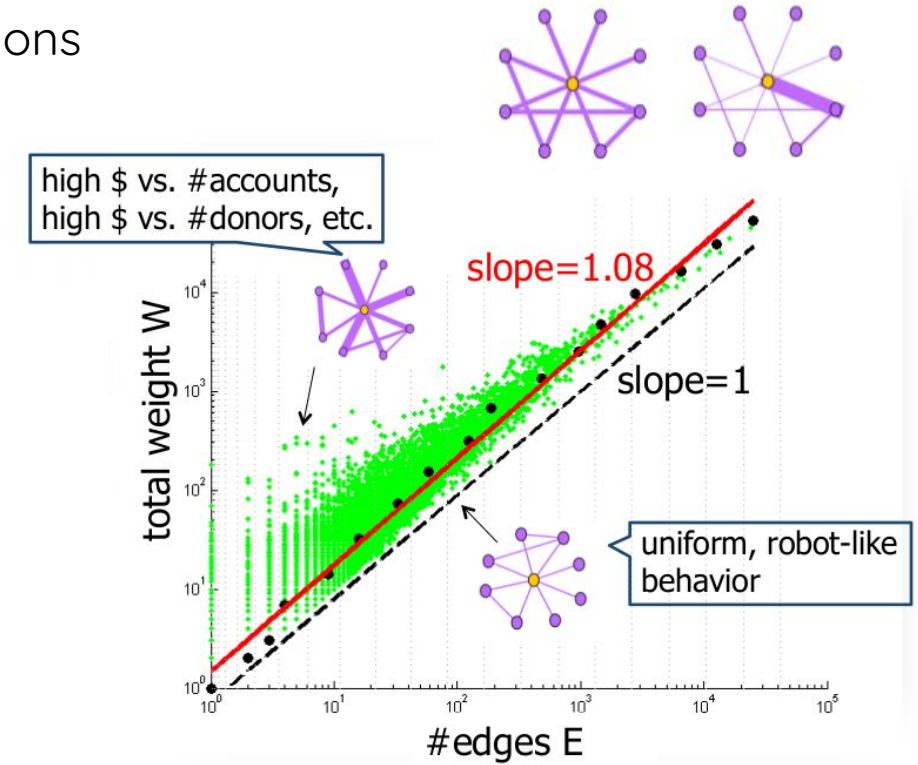


Oddball: HeavyVicinity Pattern

detects many recurrences of interactions

Egonet Weight Power Law

$W_i \propto E_i^{\alpha}, \quad 1 \leq \alpha$

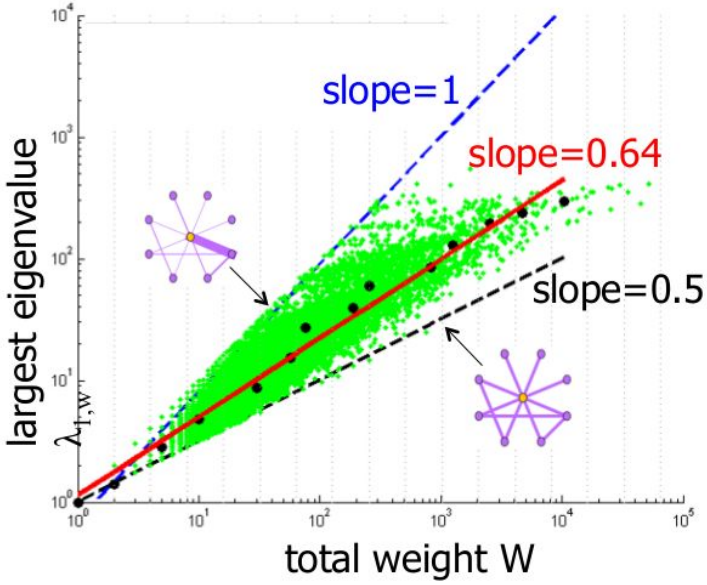


Oddball: DominantPair Pattern

detects single dominating heavy edge
(strongly connected pair)

Egonet Rank Power Law

$\lambda_{w,i} \propto W_i^{\alpha}$, $0.5 \leq \alpha \leq 1$

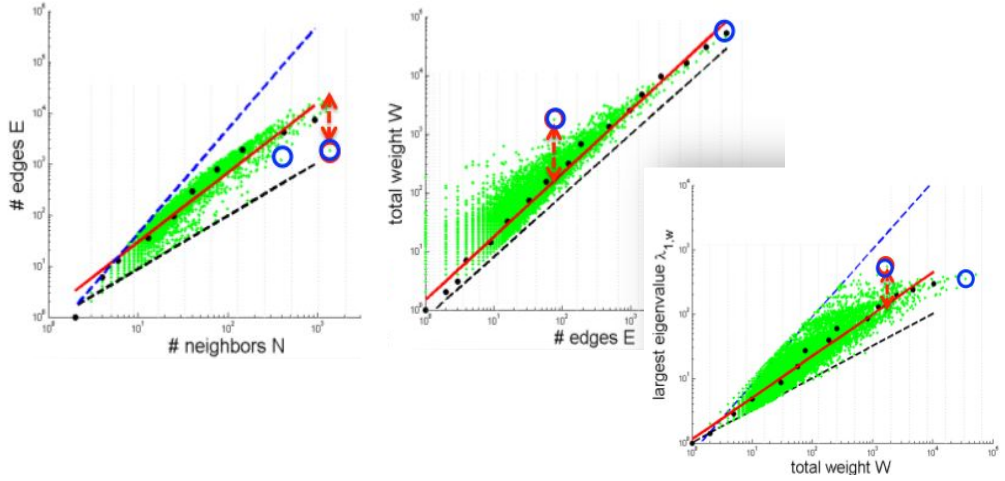


Oddball: ranking anomalies

distance to fitting line gives the anomaly score

$$\log(E_i - N_i^{\alpha} + 1)$$

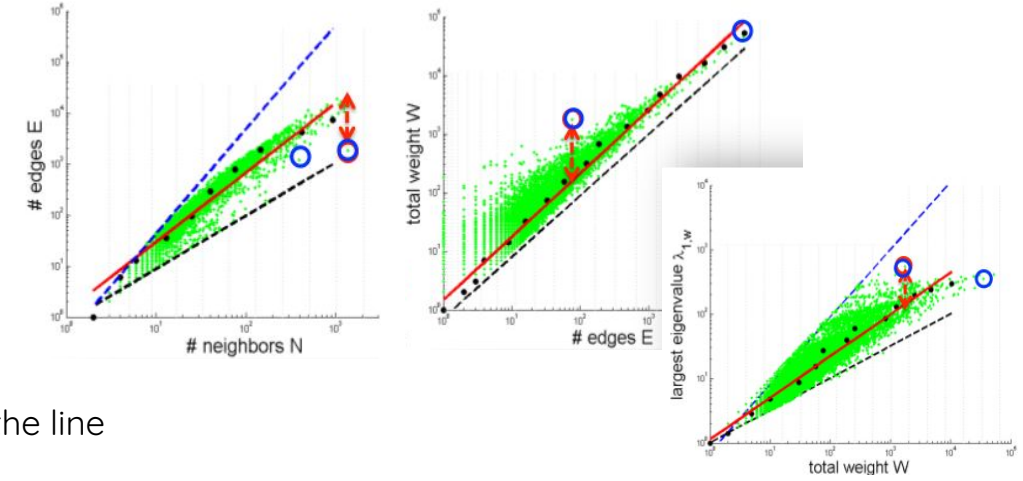
Is this enough?



Oddball: ranking anomalies

distance to fitting line gives the anomaly score

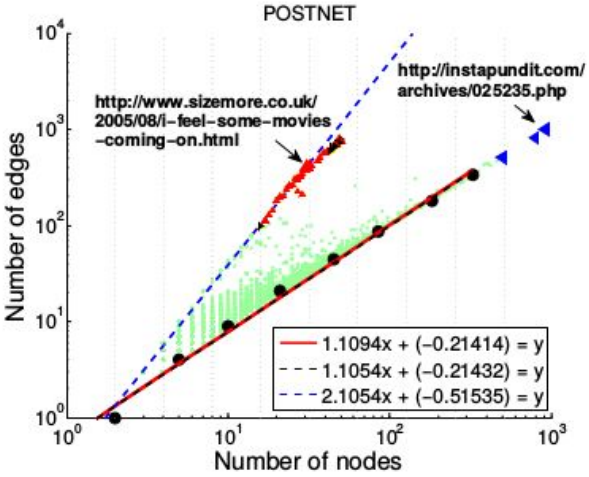
$$\log(E_i - N_i^{\alpha} + 1)$$



Is this enough?

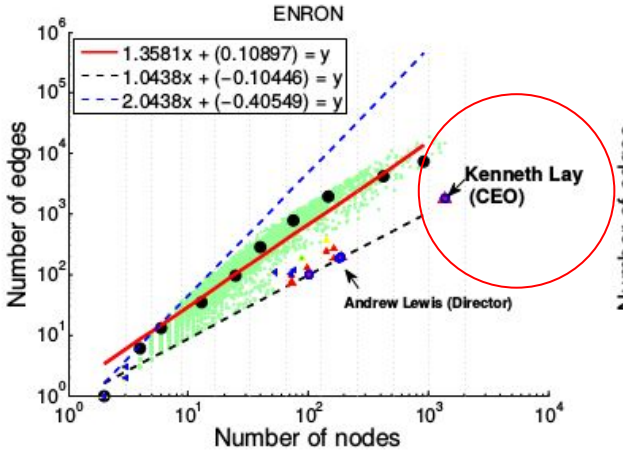
Ignores nodes that are at the extremes but on the line

Oddball: detected anomalies in CliqueStar Pattern



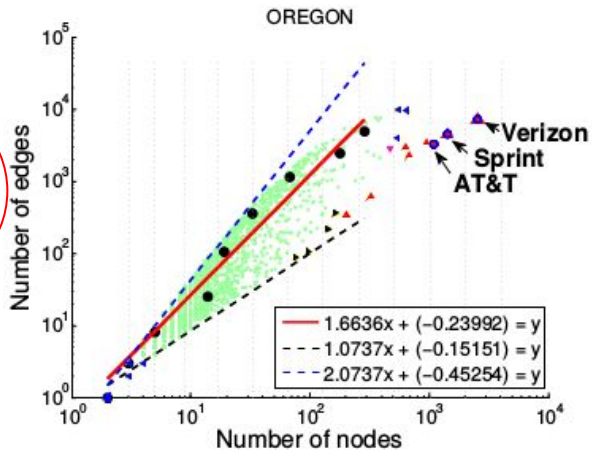
(a) *Postnet*

Network of posts based on citations



(b) *Enron*

Email associations at Enron

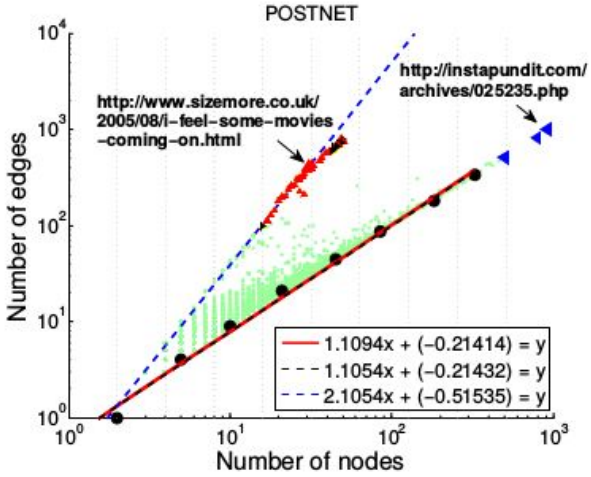


(c) *Oregon*

AS peering connections

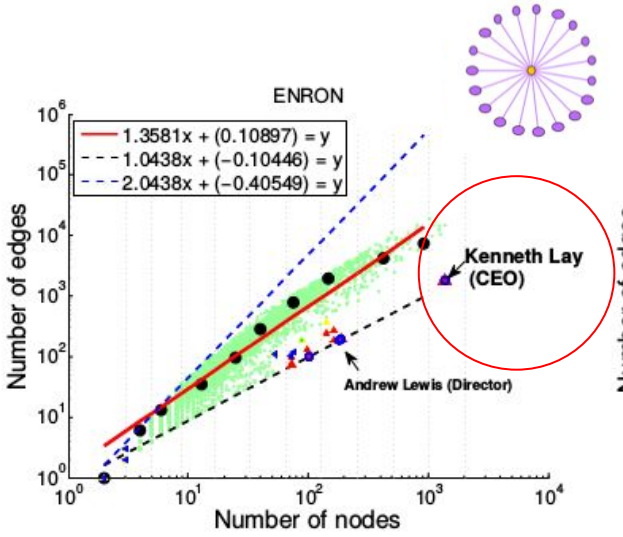
What does it mean?

Oddball: detected anomalies in CliqueStar Pattern



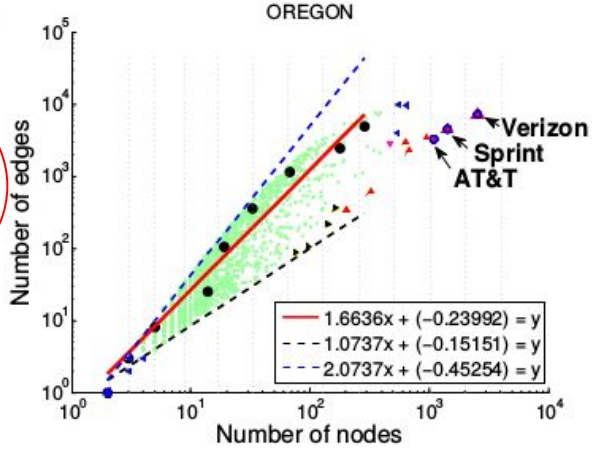
(a) *Postnet*

Network of posts based on citations



(b) *Enron*

Email associations at Enron

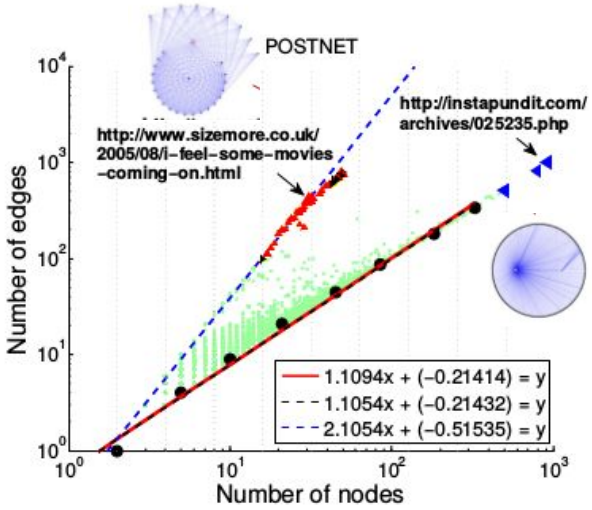


(c) *Oregon*

AS peering connections

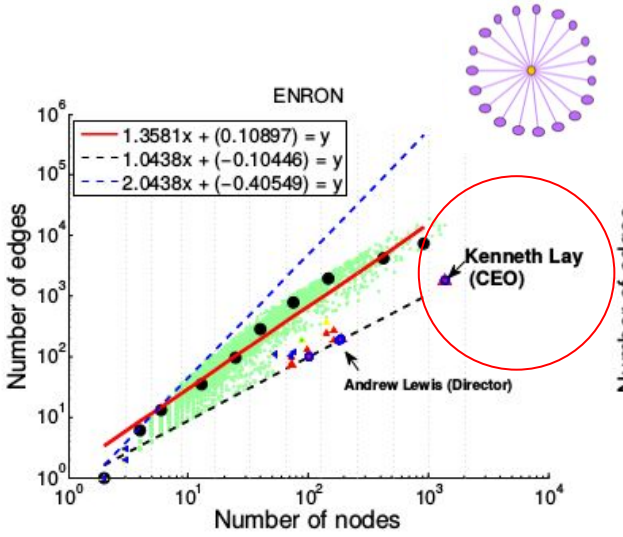
What does it mean? none of his over 1K contacts ever sent emails to each other

Oddball: detected anomalies in CliqueStar Pattern



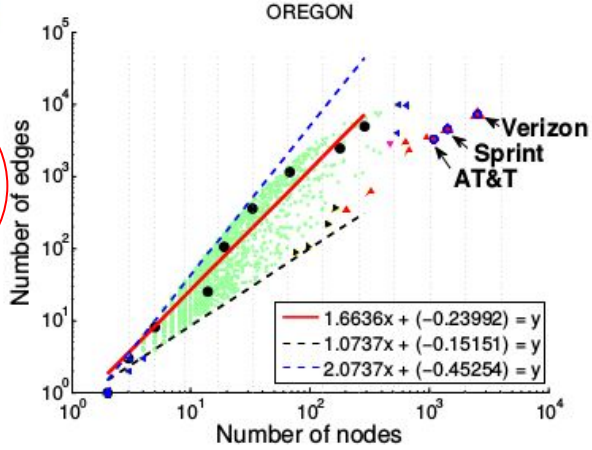
(a) *Postnet*

Network of posts based on citations



(b) *Enron*

Email associations at Enron

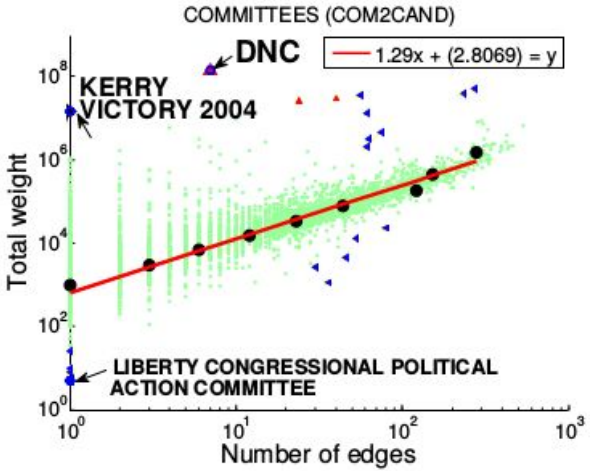


(c) *Oregon*

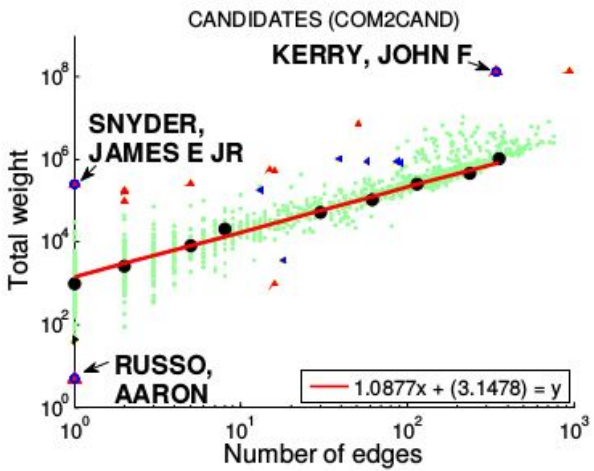
AS peering connections

What does it mean? none of his over 1K contacts ever sent emails to each other

Oddball: detected anomalies in HeavyVicinity Pattern

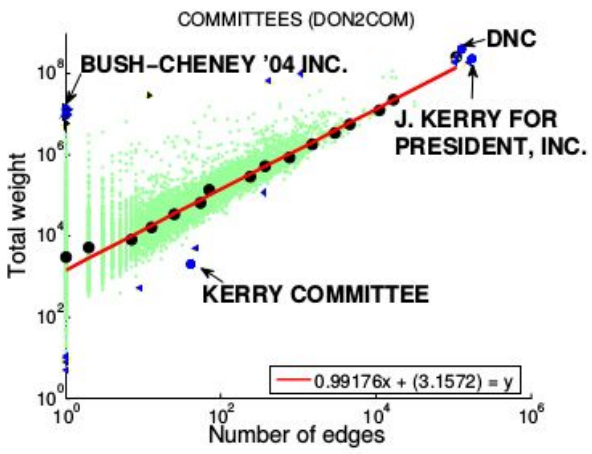


(a) *Com2Cand*



(b) *Com2Cand*

2004 US FEC Committee to Candidate donations: bipartite graph

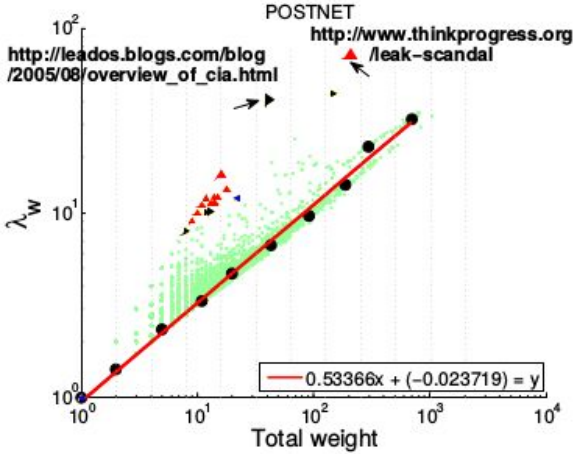


(c) *Don2Com*

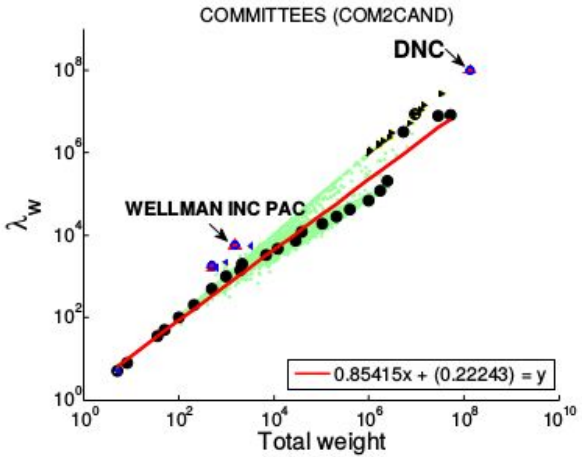
2004 US FEC Donor to Committee donations

These plots help in explaining, **interpreting** anomalies:
e.g. DNC donating a lot but to a few

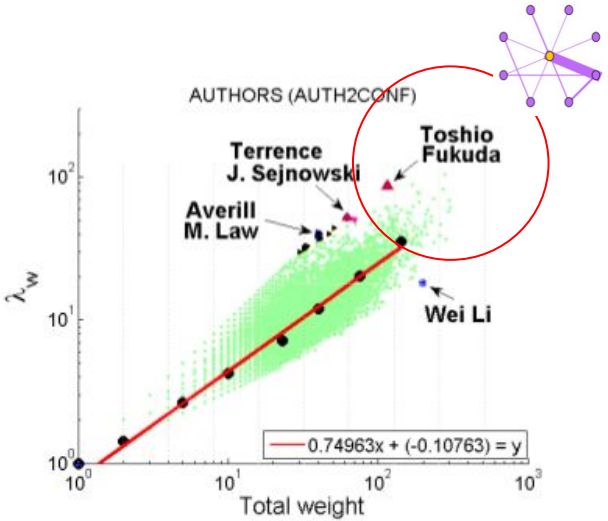
Oddball: detected anomalies in DominantPair Pattern



(a) *Postnet*



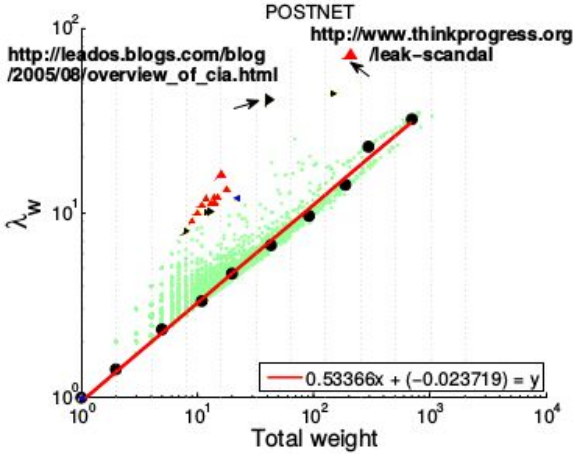
(b) *Com2Cand*



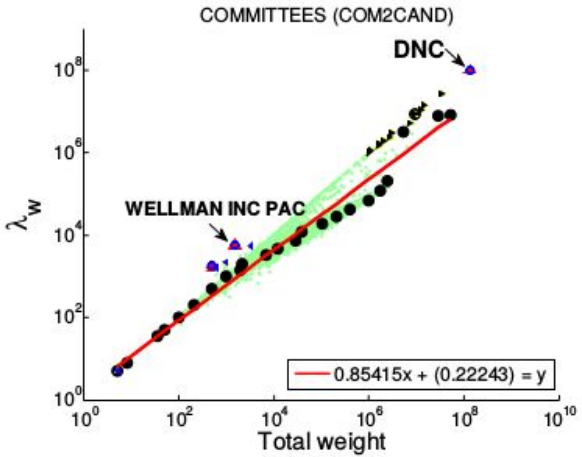
(c) *Auth2Conf*

What does it mean?

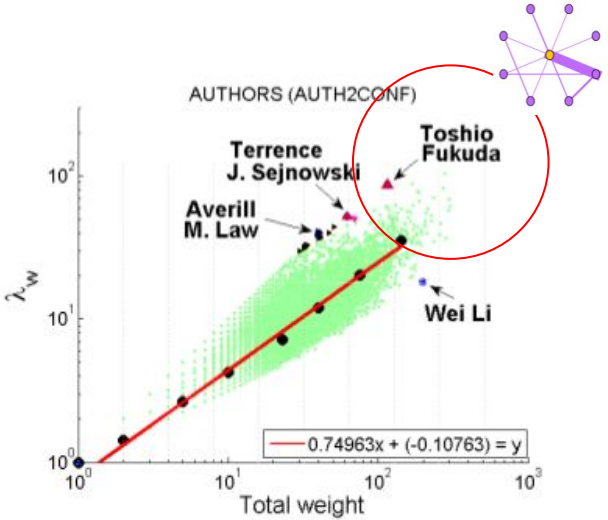
Oddball: detected anomalies in DominantPair Pattern



(a) *Postnet*



(b) *Com2Cand*



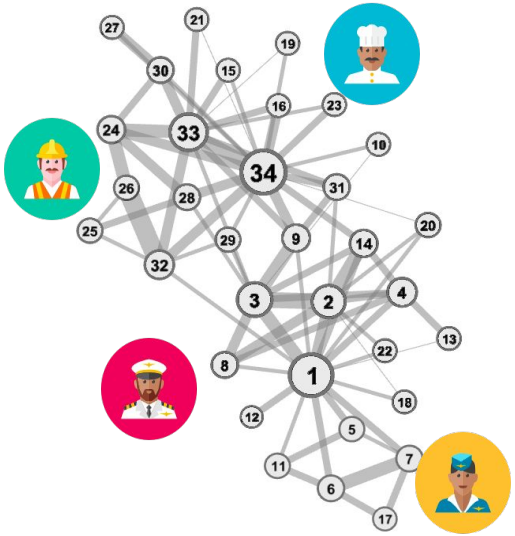
(c) *Auth2Conf*

What does it mean? Publishes mostly in one conference

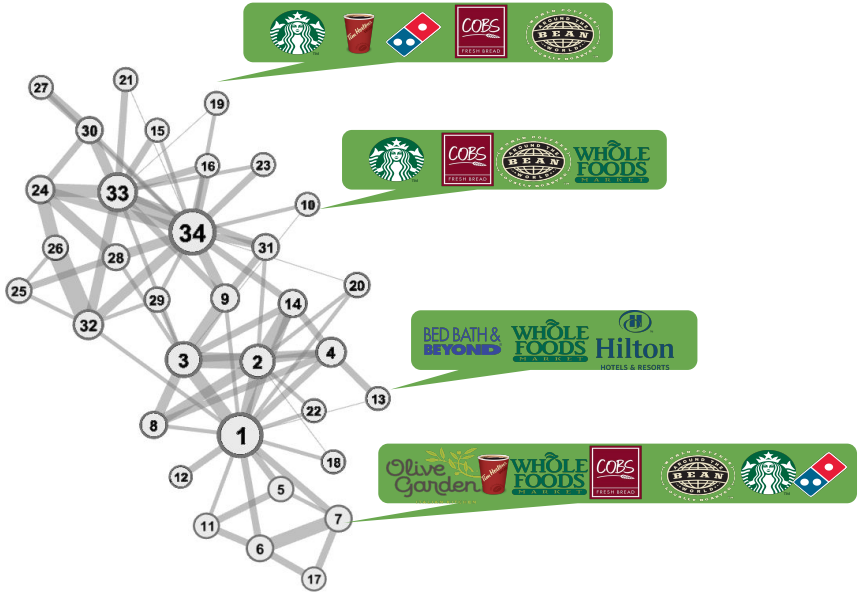
Outline

- Quick Notes
- Quick Recap of Modules
- Anomalies
  - Anomaly detection challenges and applications
  - Feature based anomaly detection (not-normal)
    - Oddball example
    - **SOAR example in attributed graphs**
    - Lookout example in explaining anomalies
  - Dense block detection (coordinated)
  - Other approaches

Anomalies in Attributed Networks

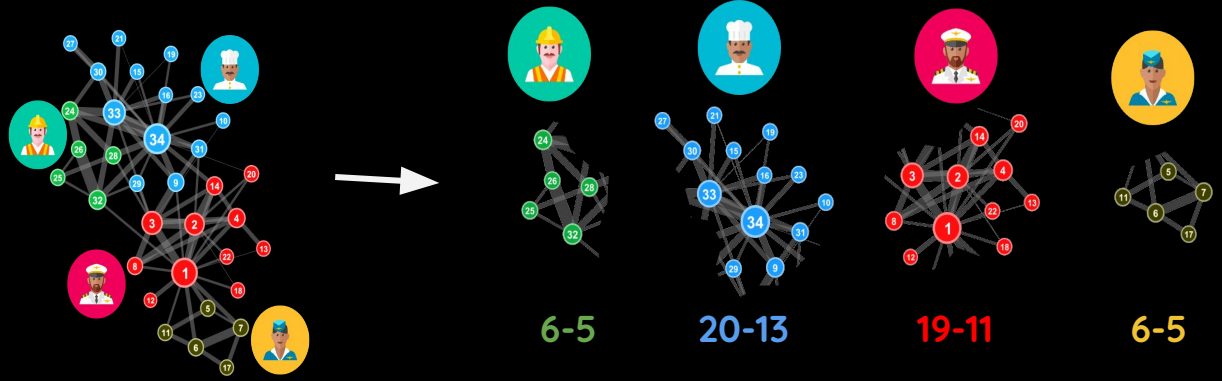


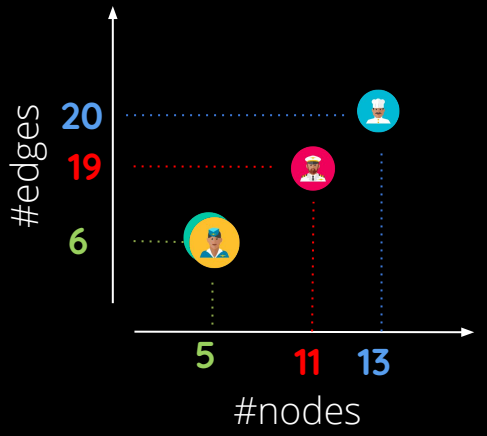
characteristics



activity & interest

Patterns in Attributed Networks



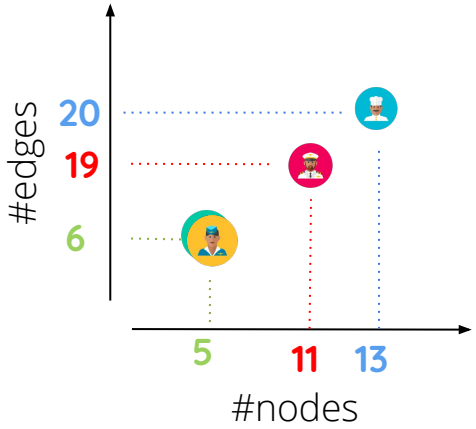


Example: #edges - #nodes

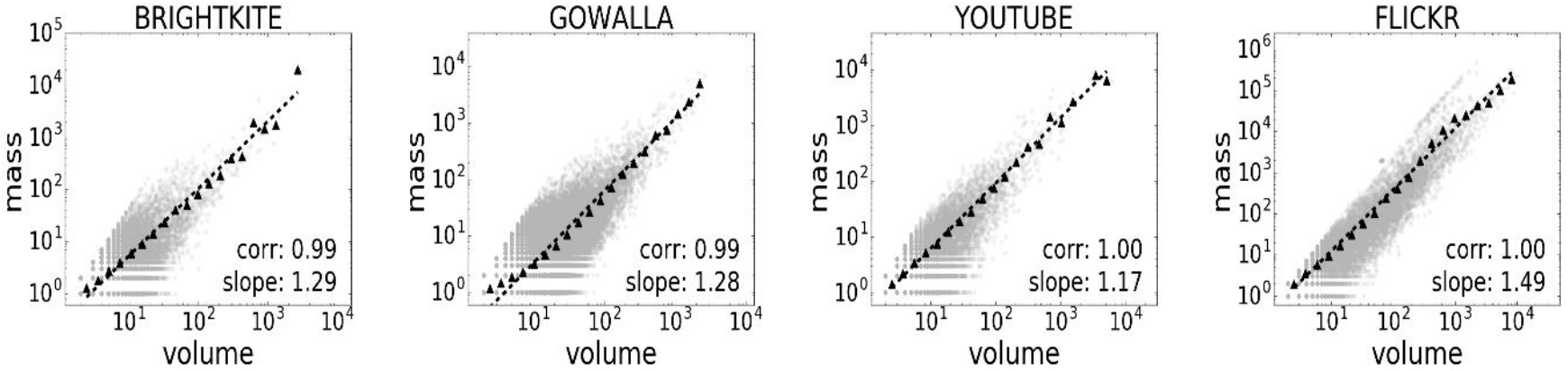
D. Eswaran, R. Rabbany, A. W. Dubrawski, C. Faloutsos, Social-Affiliation Networks: Patterns, Anomalies and a Generator, ECML/PKDD2018

Patterns in Attributed Networks

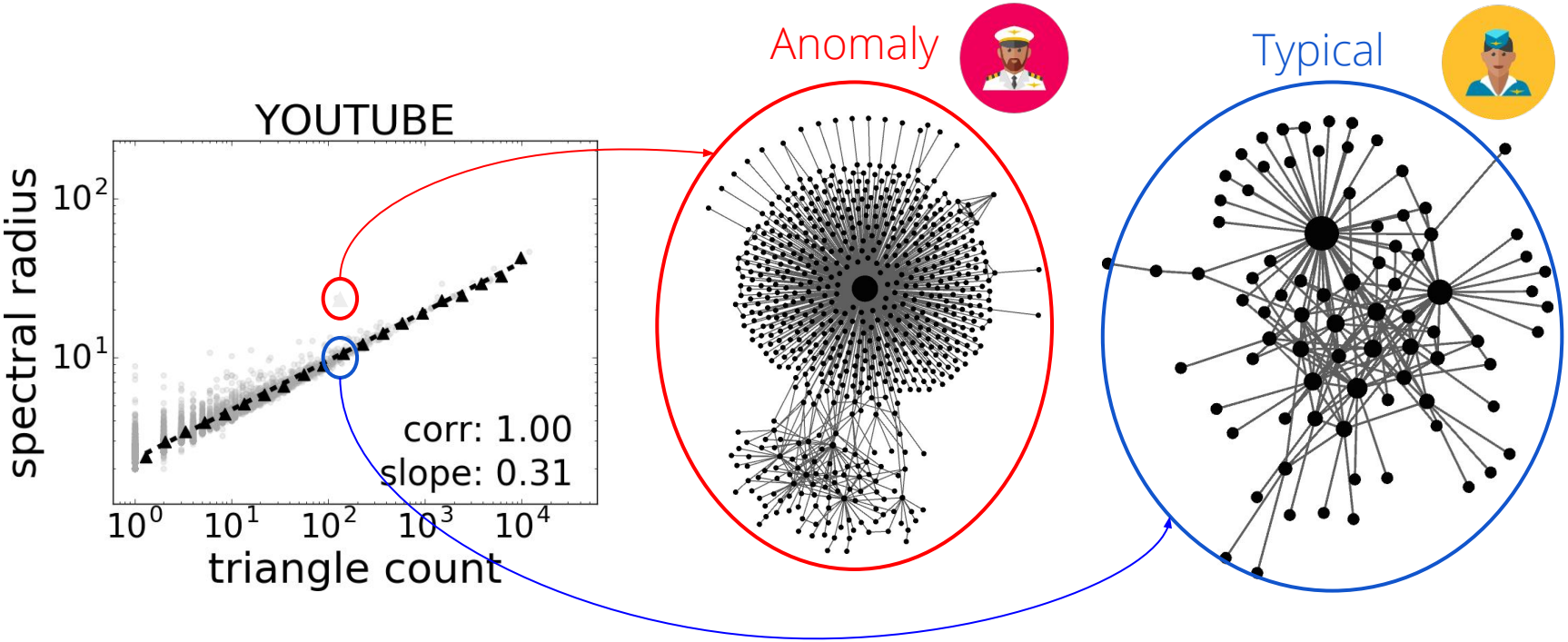
All induced subgraphs for each attribute value, all nodes sharing an attribute,
e.g. all doctors, all nurses, etc.



Example:
#edges - #nodes



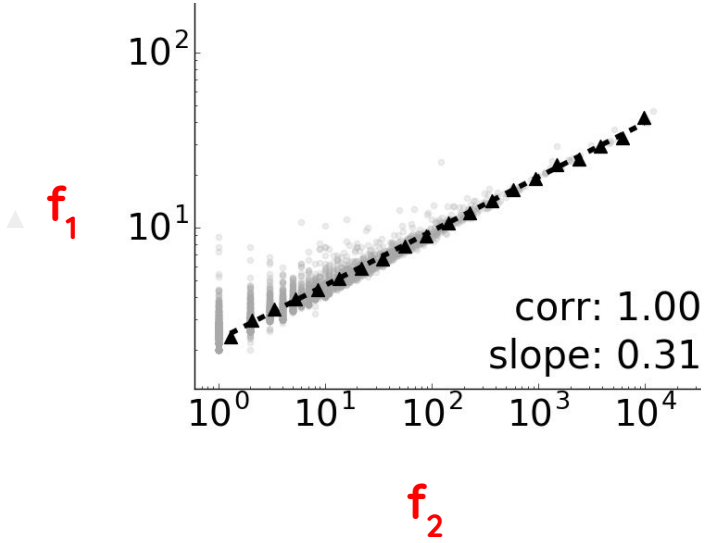
Anomalies in Attributed Networks



Outline

- Quick Notes
- Quick Recap of Modules
- Anomalies
  - Anomaly detection challenges and applications
  - Feature based anomaly detection (not-normal)
    - Oddball example
    - SOAR example in attributed graphs
    - **Lookout example in explaining anomalies**
  - Dense block detection (coordinated)
  - Other approaches

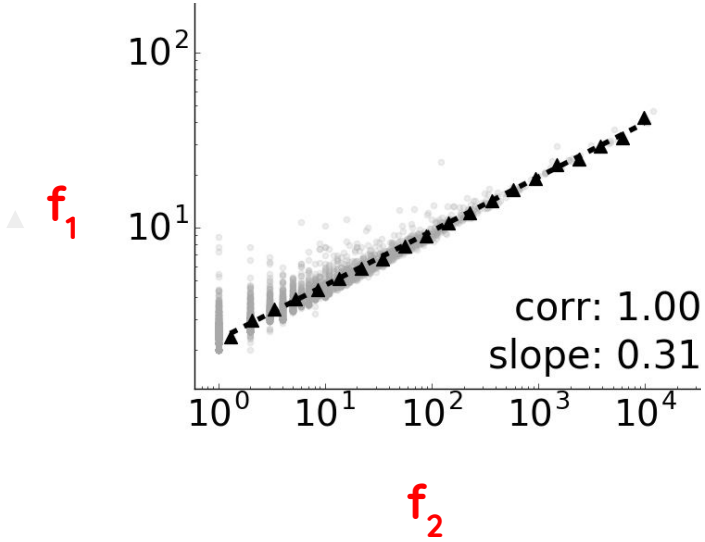
How to Explain Given Outliers?



12 measurements ⇒ how many plots?

(1) indegree and (2) outdegree for the number of unique in- and out- neighbors of every node,
(3) inweight-v and (4) outweight-v for the total weight of in- and out- edges incident on each node,
(5) inweight-r and (6) outweight-r for the count of in- and out- edges (including repetitions) incident on each node,
(7) average-IAT, (8) IAT-variance, (9-11) minimum-IAT, median-IAT, and maximum-IAT to capture various statistics of inter-arrival time (IAT) between edges and finally, (**12**) lifetime- for the time gap between the first and the last edge

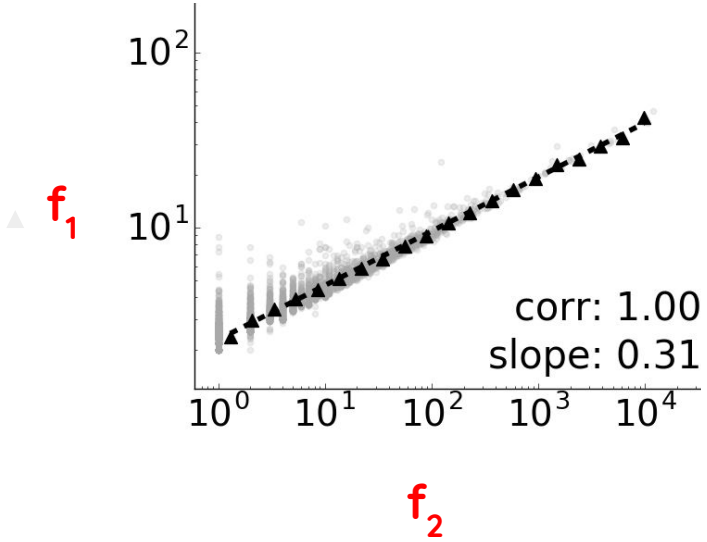
How to Explain Given Outliers?



12 measurements ⇒ how many plots? 66 x **D**

(1) indegree and (2) outdegree for the number of unique in- and out- neighbors of every node,
(3) inweight-v and (4) outweight-v for the total weight of in- and out- edges incident on each node,
(5) inweight-r and (6) outweight-r for the count of in- and out- edges (including repetitions) incident on each node,
(7) average-IAT, (8) IAT-variance, (9-11) minimum-IAT, median-IAT, and maximum-IAT to capture various statistics of inter-arrival time (IAT) between edges and finally, (**12**) lifetime- for the time gap between the first and the last edge

LookOut for Pictorial Explanation



Find useful plots automatically:
Gupta N, Eswaran D, Shah N, Akoglu L, Faloutsos C. Beyond Outlier Detection: LOOKOUT for Pictorial Explanation. InJoint European Conference on Machine Learning and Knowledge Discovery in Databases 2018 Sep 10 (pp. 122-138). Springer, Cham.

(1) indegree and (2) outdegree for the number of unique in- and out- neighbors of every node,
(3) inweight-v and (4) outweight-v for the total weight of in- and out- edges incident on each node,
(5) inweight-r and (6) outweight-r for the count of in- and out- edges (including repetitions) incident on each node,
(7) average-IAT, (8) IAT-variance, (9-11) minimum-IAT, median-IAT, and maximum-IAT to capture various statistics of inter-arrival time (IAT) between edges and finally, (**12**) lifetime- for the time gap between the first and the last edge

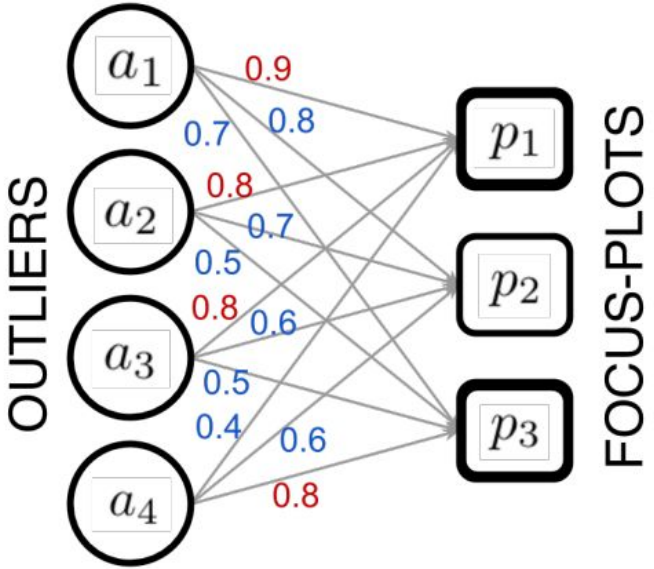
LookOut for Pictorial Explanation

Maximize the total maximum outlier score of each outlier amongst the selected plots

$$\underset{\mathcal{S}\subseteq\mathcal{P},|\mathcal{S}|=b}{\text{maximize}} \quad f(\mathcal{S}) = \sum_{a_i \in \mathcal{A}} \max_{p_j \in \mathcal{S}} s_{i,j}$$

$s_{i,j}$ depicts the outlier score that a_i received from p_j

Np-hard but has approximation with guarantees

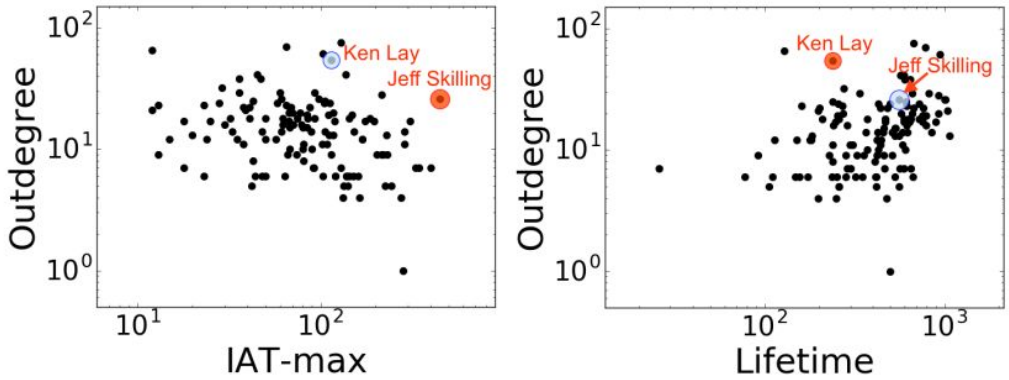


LookOut for Pictorial Explanation

LookOut explains Enron founder/CEO “Ken Lay” and COO “Jeff Skilling” by two “pair plots” in which they are most salient. Compared to traditional ranked list output (left), LookOut produces simpler, more interpretable explanations (right)

Name	Score
Skilling, Jeff	0.893
Lay, Kenneth	0.761
Fastow, Andrew	0.442
Mark, Rebecca	0.429
Smith, John	0.331
Cooper, Stephen	0.308
Tomson, Mary	0.232
...	...

(a) traditional



(b) proposed

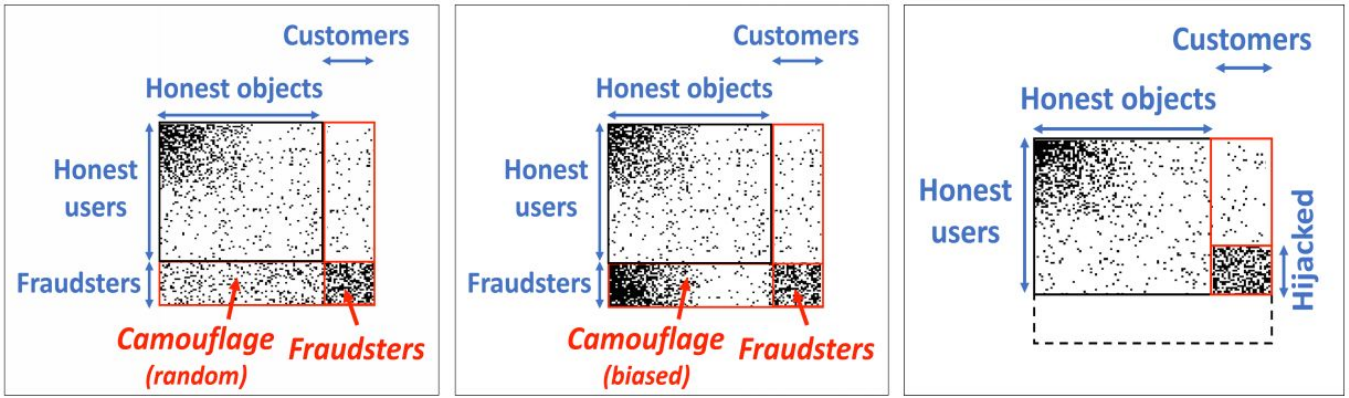
Outline

- Quick Notes
- Quick Recap of Modules
- Anomalies
  - Anomaly detection challenges and applications
  - Feature based anomaly detection (not-normal)
    - Oddball example
    - SOAR example in attributed graphs
    - Lookout example in explaining anomalies
  - **Dense block detection (coordinated)**
  - Other approaches

Anomalies as dense blocks

"How can we detect if a politician has purchased fake followers on Twitter, or if a product's reviews on Amazon are genuine? fraudsters add many edges, creating unusually large and dense regions in the adjacency matrix of the graph, smart fraudsters will also try to look normal, by adding links to popular items/idols"

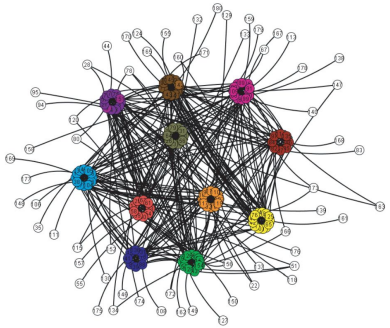
Hooi B, Song HA, Beutel A, Shah N, Shin K, Faloutsos C. **Fraudar**: Bounding graph fraud in the face of camouflage, KDD 2016

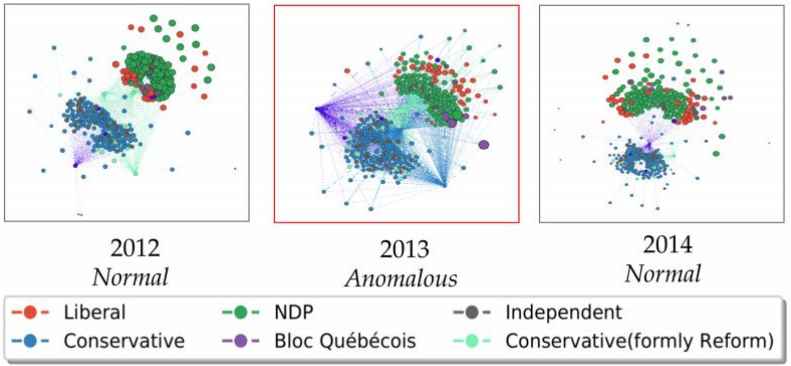


(a) Attacks with random camouflage (b) Attacks with biased camouflage (c) Hijacked accounts

Other approaches

- Minimum Description Length
  - Rare things that take longer to explain
- Community based
  - Outliers that do not belong to any module
- Supervised, semi supervised
  - Classification using marked spams, scams
- Embedding based
  - Instead of handcrafted anomalies
- Dynamic Graphs
  - Incremental, online, streams
  - Detect events and change points





Huang S, Hitti Y, Rabusseau G, Rabbany R. Laplacian Change Point Detection for Dynamic Graphs. KDD 2020