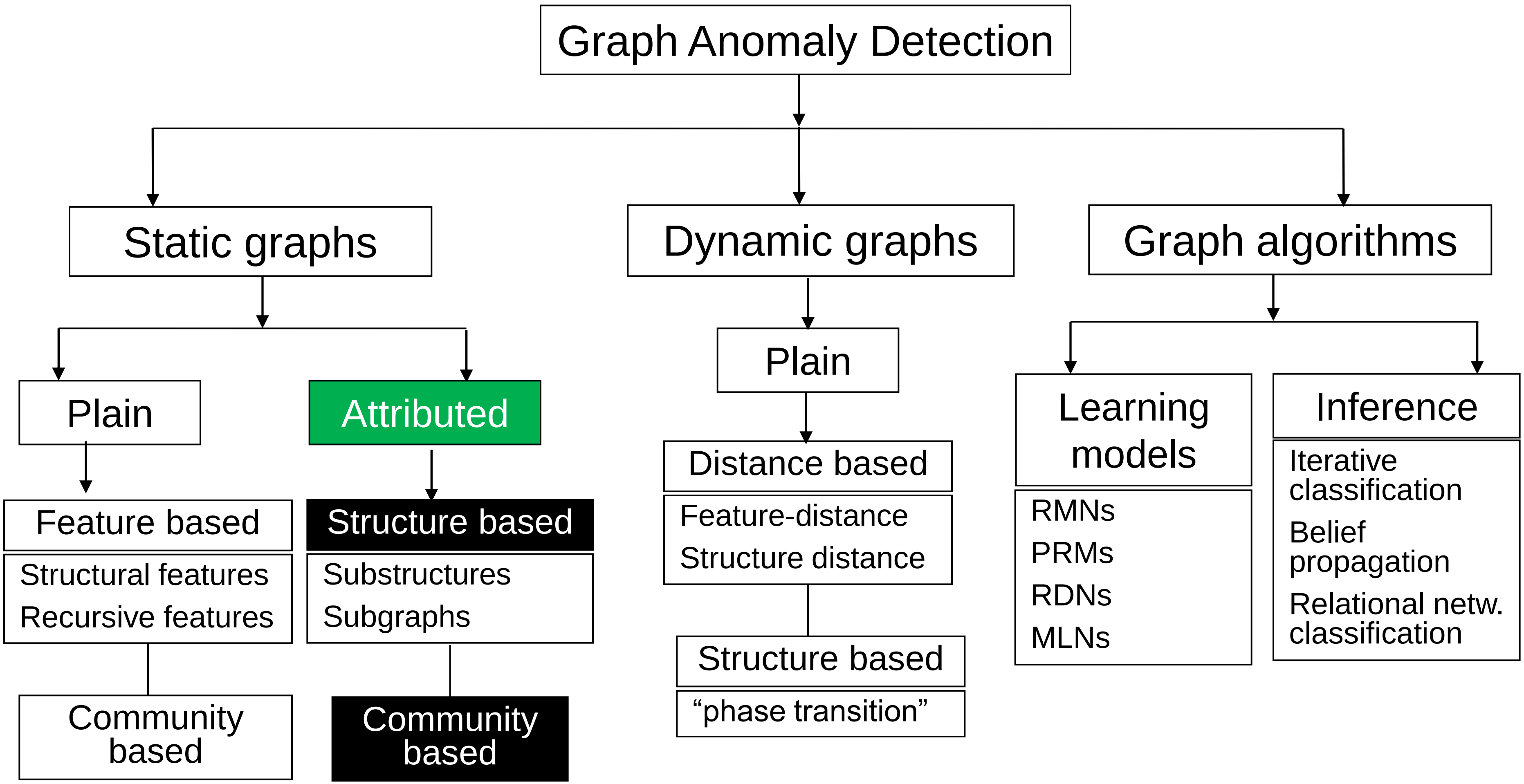
Taxonomy



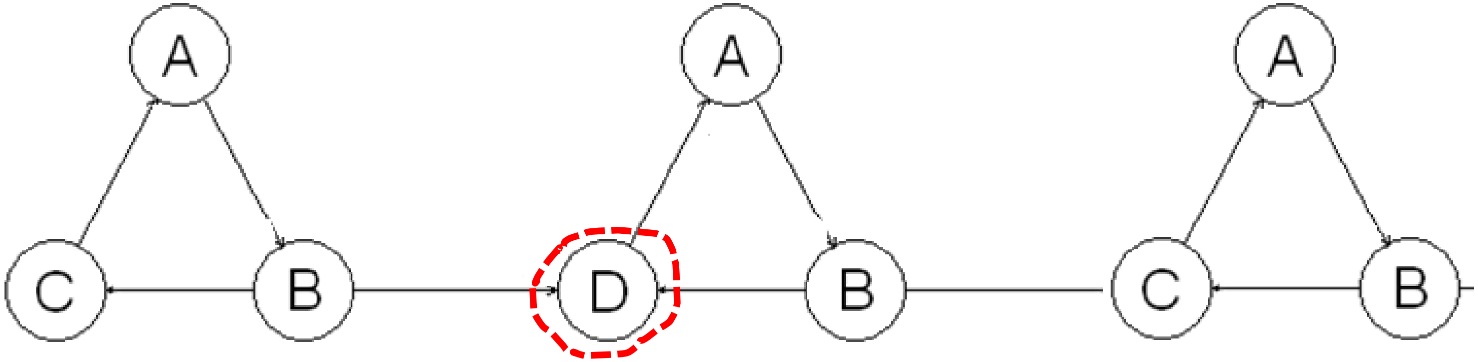
Graph Anomaly Detection
Static graphs
Dynamic graphs
Graph algorithms
Plain
Attributed
Feature based
Structural features
Recursive features
Community based
Structure based
Substructures
Subgraphs
Community based
Plain
Distance based
Feature-distance
Structure distance
Structure based
“phase transition”
Learning models
RMNs
PRMs
RDNs
MLNs
Inference
Iterative classification
Belief propagation
Relational netw. classification

Noble & Cook. '03

Anomalies in labeled graphs

- Problem:

Q1. Given a graph in which nodes and edges contain (non-unique) labels, what are unusual substructures?



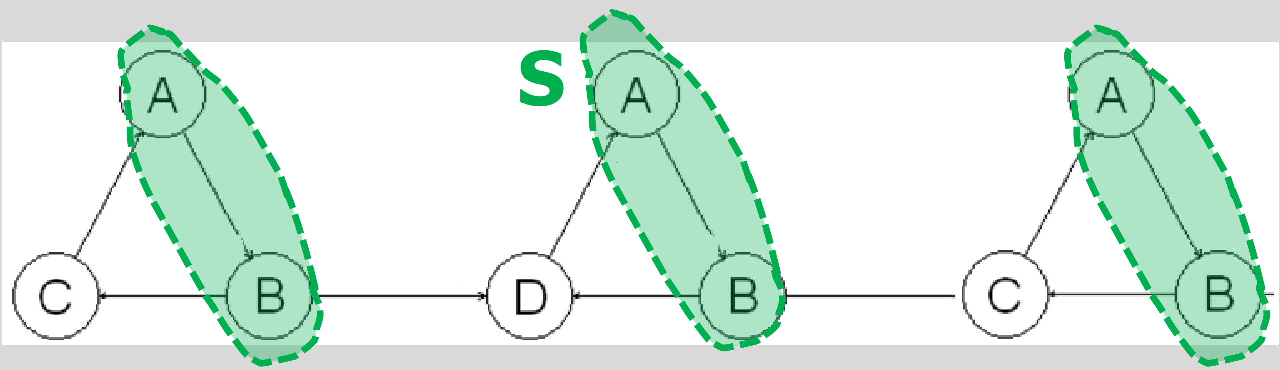
Background

- Subdue*: An algorithm for detecting repetitive patterns (substructures) within graphs.
- Substructure: A connected subgraph of the overall graph.
- Compressing a graph: Replacing each instance of the substructure with a new vertex representing that substructure.
- Description Length (DL): Number of bits needed to encode a piece of data

* http://ailab.wsu.edu/subdue/

Background

- Subdue uses the following heuristic:
  - The best substructure is the one that **minimizes** **F1(S,G) = DL(G | S) + DL(S)**
    - G: Entire graph, S: The substructure,
    - DL(G|S) is the DL of G after compressing it using S,
    - DL(S) is the description length of the substructure.



- Iterations after compressing at each step

Background

Given database D and set of models for D, **Minimum Description Length** selects model M that minimizes

$$\underbrace{L(M)} + \underbrace{L(D|M)}$$

length in bits: description of **model** M

length in bits: **data**, encoded by M

⇩

a_1x+a_0

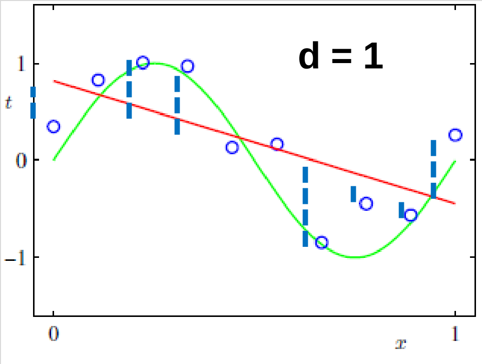
vs.

$a_9x^9+\ldots+a_1x+a_0$

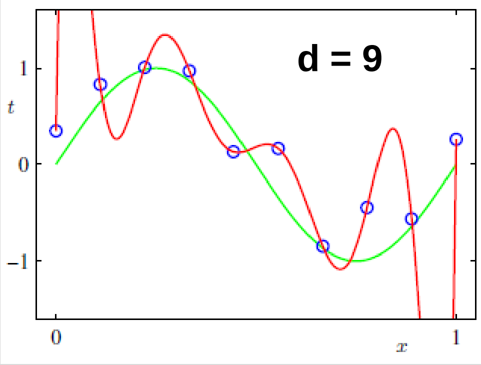
⇩

deltas

{}



vs.



Bishop: PR&ML

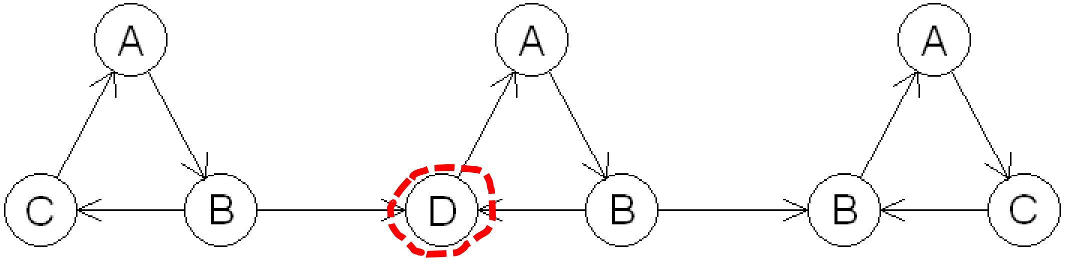
1) Anomalous Substructures

- Main idea: anomalies (by def.) occur infrequently, they are roughly opposite to “best substructures”
  - Find substructures **S** that maximize **F1(S,G)**?
    - Nope, it flags all single nodes as anomalies!
  - Instead, find those that **minimize**

    F2(S, G) = Size(S) * Instances(S,G)
    - Approximate inverse of **F1(S,G)**
- Intuition: Larger substructures are <u>expected</u> to occur few times; the smaller the substructure, the less likely it is rare

Example

- **F2(S, G) = Size(S) * Instances(S,G)**
  - For node D, F2 = 1 * 1 = 1
  - For A→C and D→A, it is 2 * 1 = 2
  - For G (whole graph), it is 9 * 1 = 9
- Hence D is considered the most anomalous.



- Note: Usually a threshold for F2 is used and anomalies are ranked by their scores.

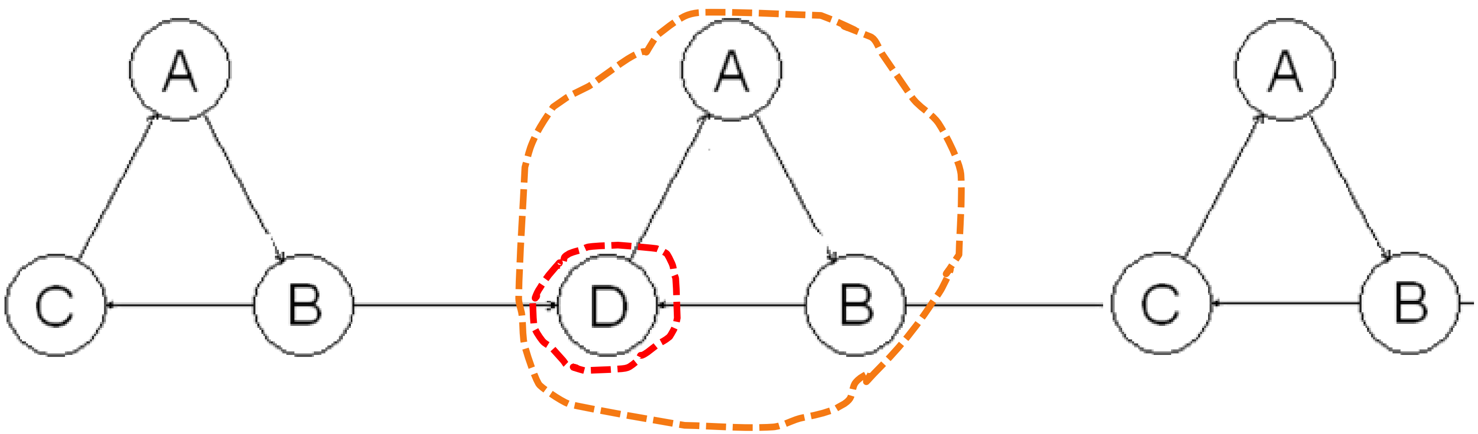
Noble & Cook. '03

Anomalies in labeled graphs

- Problem:

Q1. Given a graph in which nodes and edges contain (non-unique) labels, what are unusual substructures?

Q2. Given a set of subgraphs, what are the unusual subgraphs?



Note: assumption is anomalies are connected

2) Anomalous Subgraphs

- Main idea: subgraphs containing few common substructures are generally more anomalous
  - Define anomaly score A in [0,1]

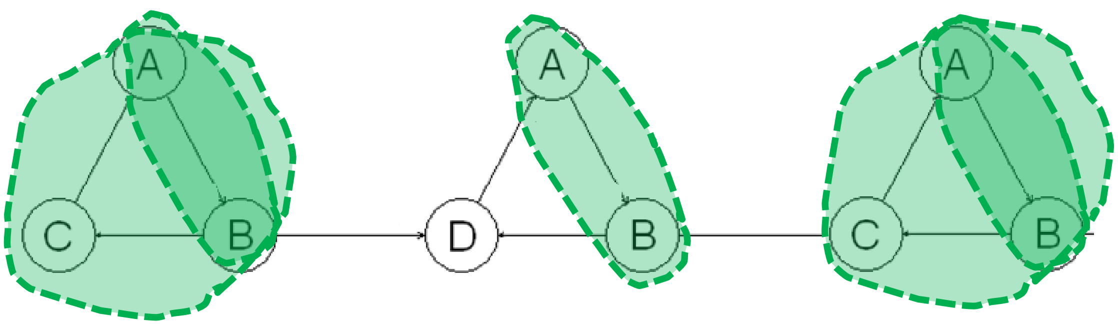
$$A = 1 - \frac{1}{n}\sum_{i=1}^{n}(n-i+1) * c_i$$

Subdue iterations

fast drop off in early iterations

fraction compressed at *ith* iteration

$$\boxed{\frac{DL_{i-1}(G) - DL_i(G)}{DL_0(G)}}$$



Experiments

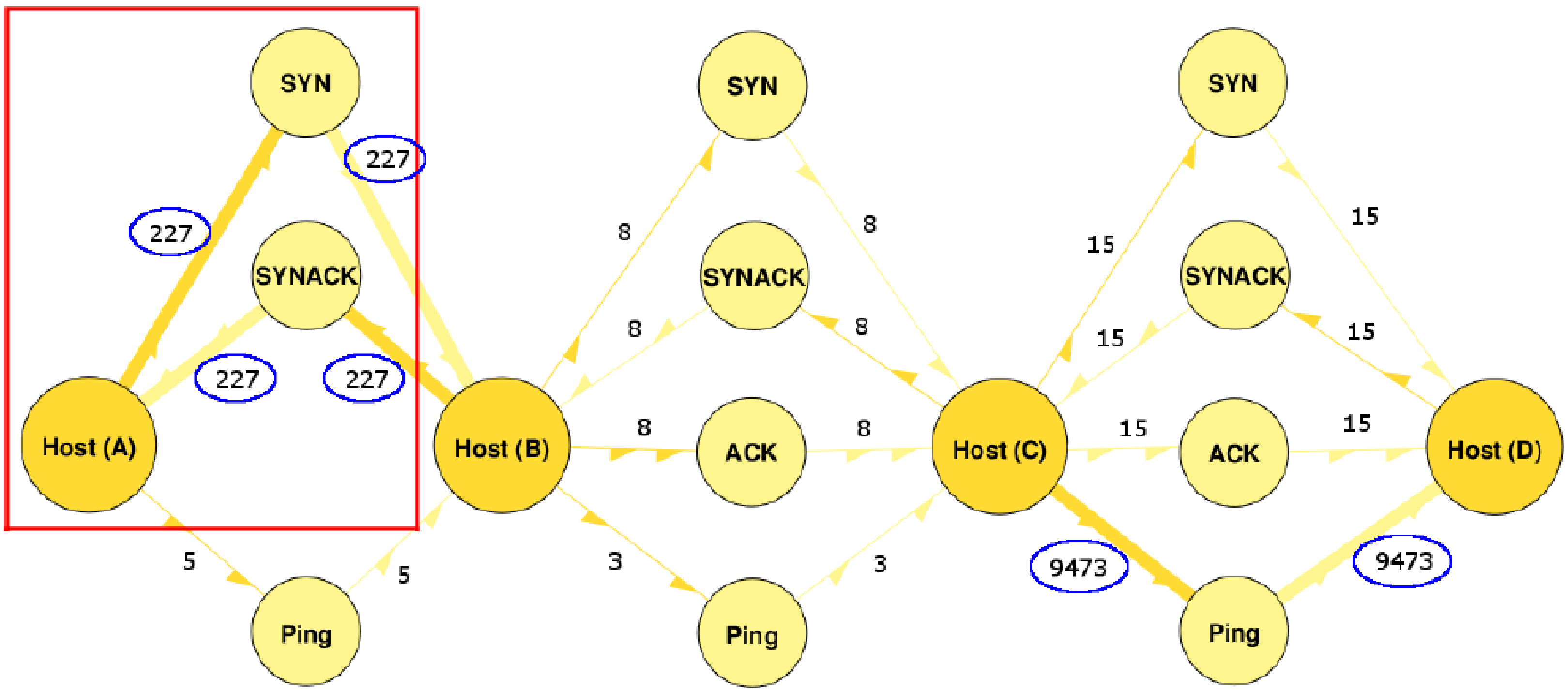
- Data: 1999 KDD Cup Network Intrusion
  - Ground truth: connection records, "normal" or attack (37 types), 41 features of connection (duration, protocol type, number of bytes, etc.)
  - Each individual test involved 50 records of which only one is of a particular attack type.
- Use Subdue to find anomalous substructures
  - Prune all subgraphs with size>3, F2>6 (arbitrary)

Davis et al. '11

Anomalies with numeric labels

- How about **numeric** labels?
  - Noble & Cook work with **categorical** labels

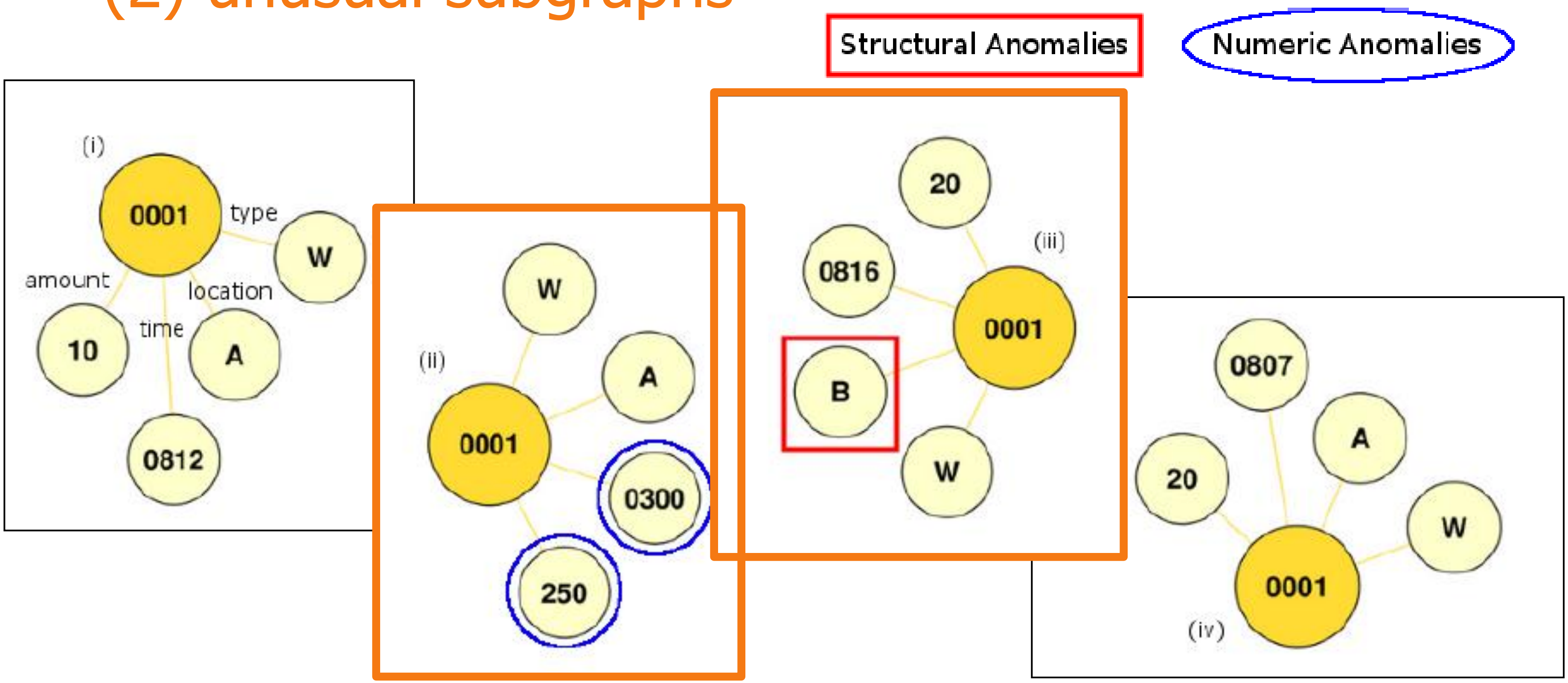
(1) unusual substructures



Anomalies with numeric labels

- How about **numeric** labels?
  - Noble & Cook work with **categorical** labels

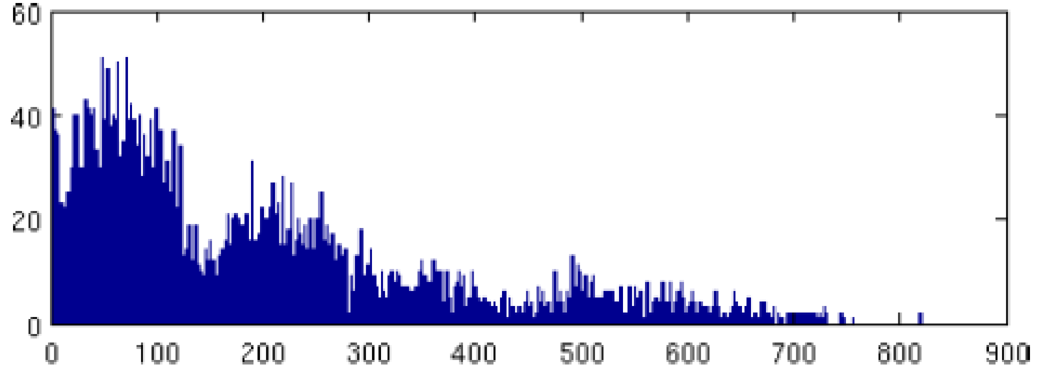
(2) unusual subgraphs



Anomalies with numeric labels

- Main idea (discretization):
  - assign categoric label q_0 to "normal" values, and
  - "outlierness" score q_i to all others i

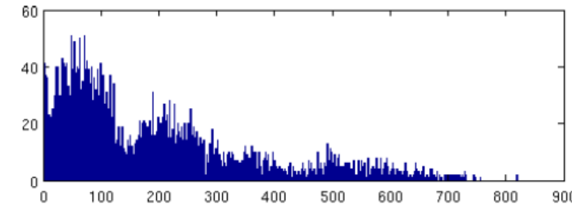
- Example: empirical distribution of a label

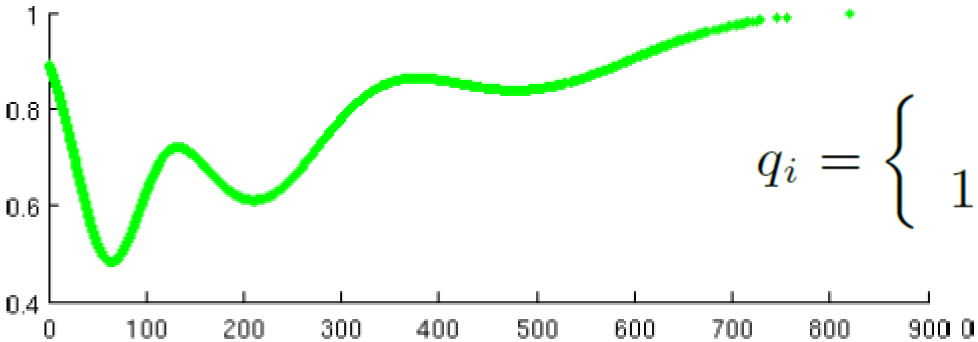


- Several "outlierness" scores (pdf-fitting, kNN, LOF, clustering-based)

Discretization

- Model fitting (GMM)

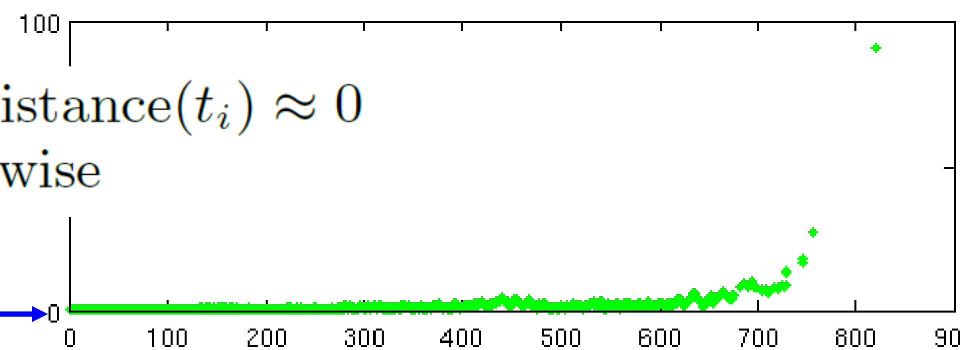




$$q_i = \begin{cases} q_0 & \text{if } 1 - P(t_i) < q_a \\ 1 - P(t_i) & \text{otherwise} \end{cases}$$

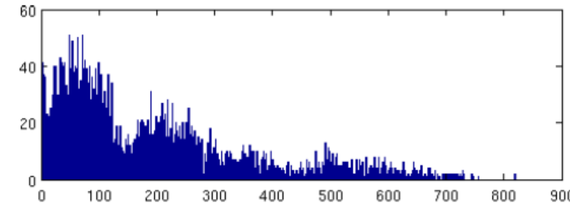
- kNN distance

$$q_i = \begin{cases} q_0 & \text{if k-distance}(t_i) \approx 0 \\ \text{k-distance}(t_i) & \text{otherwise} \end{cases}$$

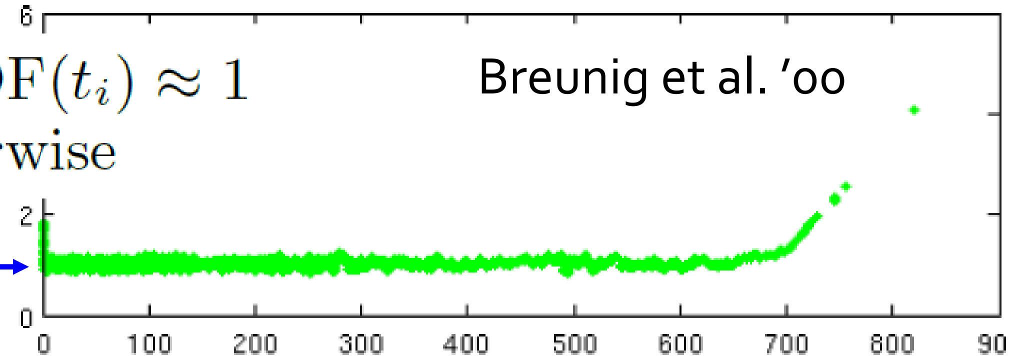


Discretization

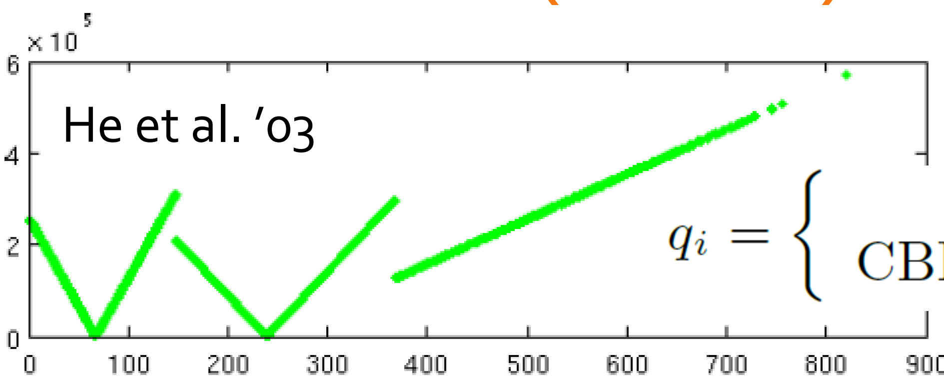
- Density outlier score (LOF)



$$q_i = \begin{cases} q_0 & \text{if } \mathrm{LOF}(t_i) \approx 1 \\ \mathrm{LOF}(t_i) & \text{otherwise} \end{cases}$$



- Cluster-based (CbLOF)



$$q_i = \begin{cases} q_0 & \text{if } \mathrm{CBLOF}(t_i) < q_a \\ \mathrm{CBLOF}(t_i) & \text{otherwise} \end{cases}$$

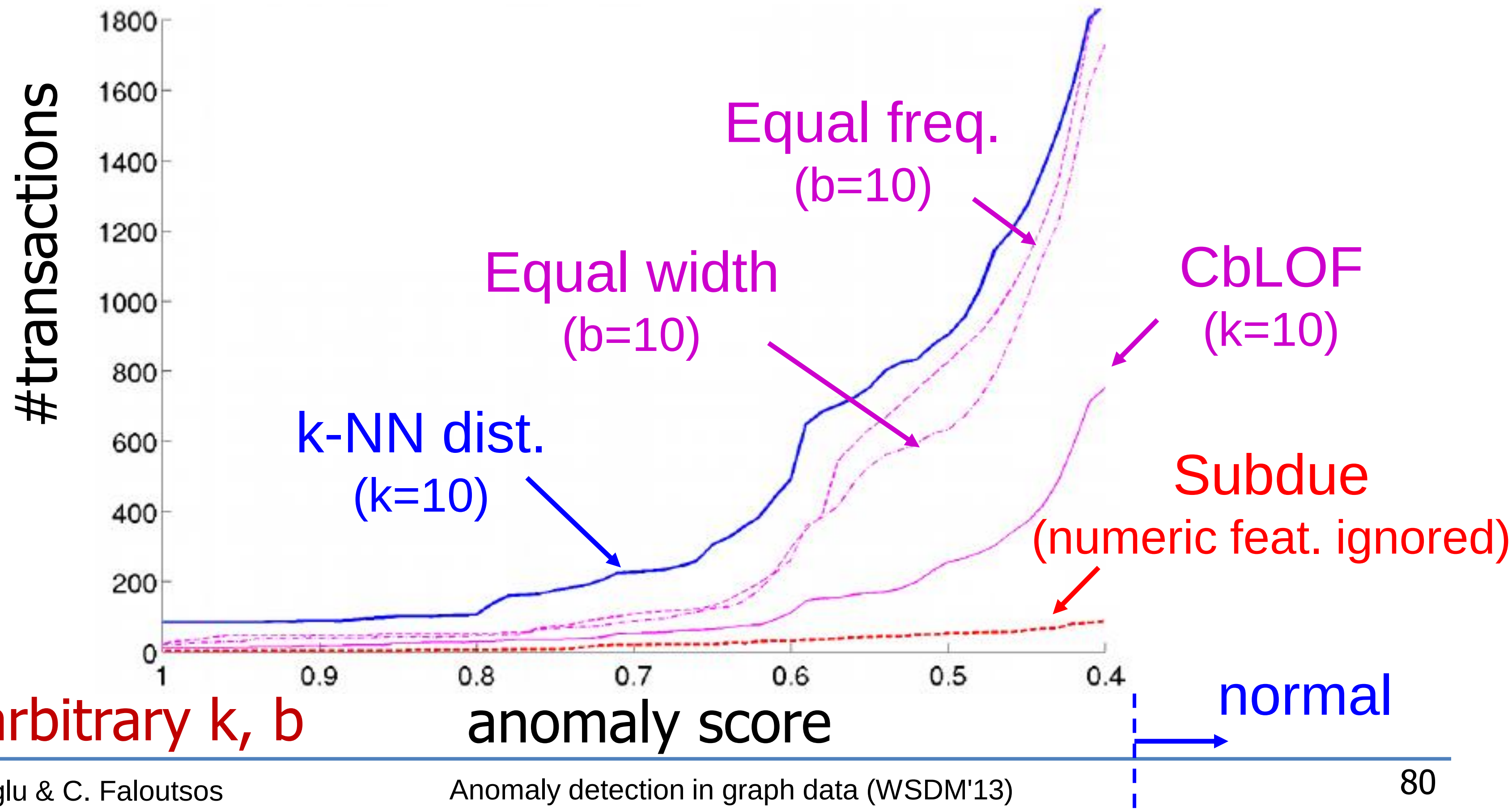
distance to closest “large” (k-means) cluster centroid

Discretization

- Other possible discretization techniques
  - SAX (**S**ymbolic **A**ggregate appro**X**imation)
    - http://www.cs.ucr.edu/~eamonn/SAX.htm
  - MDL-binning
    - P. Kontkanen and P. Myllymäki. *MDL histogram density estimation*. In AISTAT, 2007.
  - Minimum entropy discretization
    - U.M. Fayyad and K.B. Irani. *Multi-interval discretization of continuous-valued attributes for classification learning*. In Proc. IJCAI, 1989.
  - Logarithmic binning
    - especially for skewed distributions

Experiment

- Data: Access card transaction graphs
  - node: door sensor, edge (u,w): movement u→w, weight(u,w): time u→w (only numeric attribute)



* arbitrary k, b

Anomalies in labeled graphs

- Problem:

Q1. **Given** a graph in which nodes and edges contain (non-unique) labels, how to **find** substructures that are very similar to, though not the same as, a normative substructure? ("best substructure" as for Subdue)*

- Intuition:

"The more successful money-laundering apparatus is in imitating the patterns and behavior of legitimate transactions, the less the likelihood of it being exposed."

– United Nations Office on Drugs and Crime

Formal definition

- Given graph *G* with a normative substructure *S*, a substructure *S'* is anomalous if difference *d* between *S* and *S'* satisfies $0 < d <= X$, where *X* is a (user-defined) threshold and *d* is a measure of the unexpected structural difference.
- Assumptions
  - Majority of *G* consists of a normative pattern, and no more than X% of it is altered in an anomaly.
  - Anomalies consist of one or more modifications, insertions or deletions.
  - Normative pattern is connected.

Three Types of Anomalies

1) GBAD-MDL (Minimum Descriptive Length): anomalous modifications

2) GBAD-P (Probability): anomalous insertions

3) GBAD-MPS (Maximum Partial Substructure): anomalous deletions

Note: prone to miss more than one type of anomaly

- e.g., a deletion followed by modification

1) Information Theoretic Approach

- Find normative substructure S that minimizes

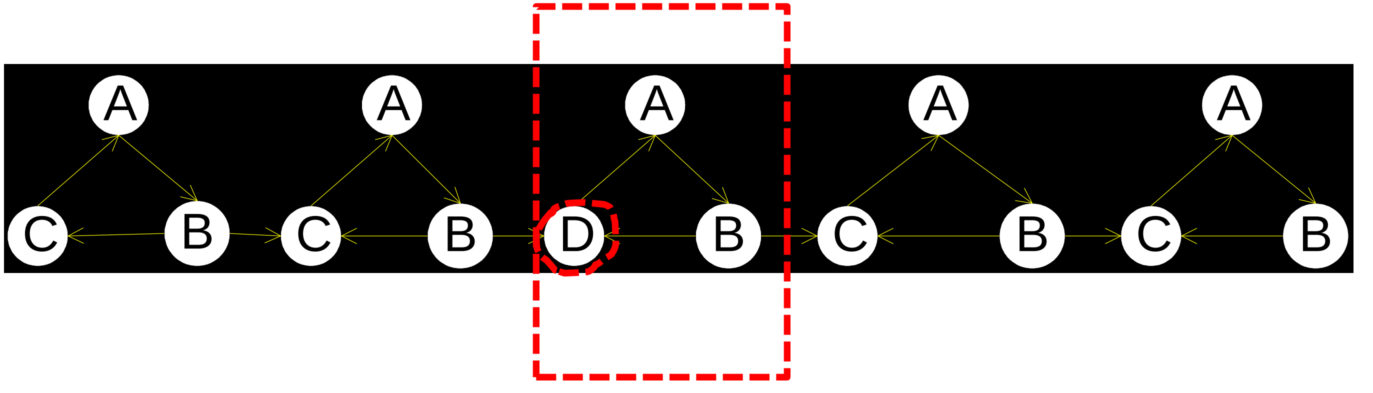
$$F(S,G) = DL(G \mid S) + DL(S)$$

- For each instance I_k of S

$$\text{anomalyScore}(I_k) = \text{freq}(I_k) * \text{matchcost}(I_k,S)$$

the lower, the more anomalous   cost to modify I_k into S

- Example:

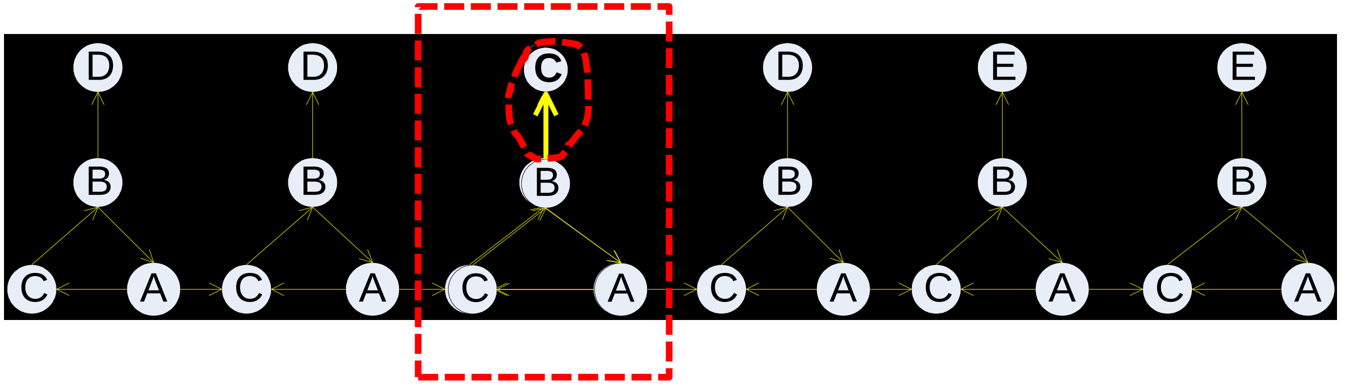


2) Probabilistic Approach

- Find normative substructure **S**
- Find extensions to **S** with lowest probability
- For each extension I_k of **S**

$$\text{anomalyScore}(I_k) = \frac{\text{number of instances of } I_k}{\text{all instances } I_n \text{ with a unique extension}}$$

- Example:



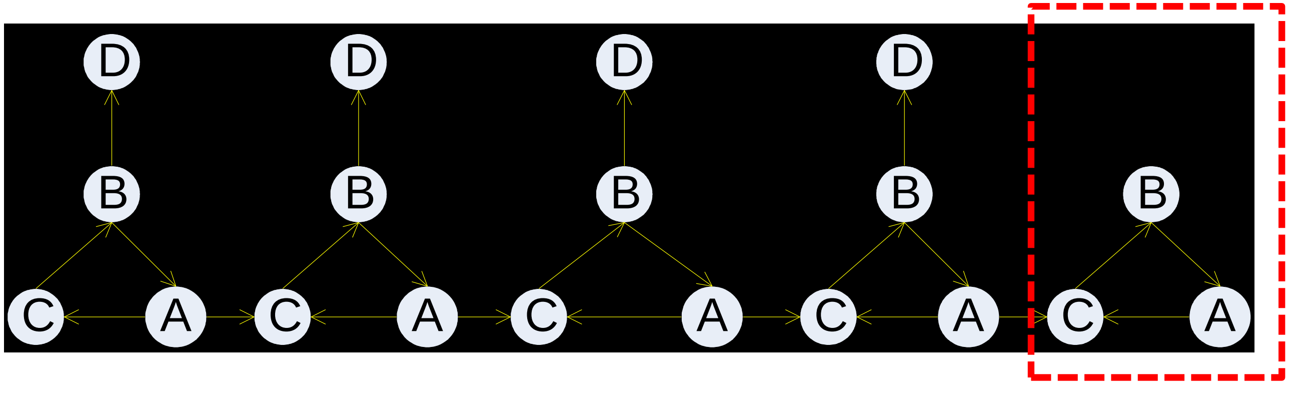
3) Maximum Partial Substructure Approach

- Find normative substructure **S**
- Find "ancestral" substructures $S_n \subseteq S$ that are missing various edges and vertices.
- For each instance $\boldsymbol{I_k}$ of $\mathbf{S}_{\boldsymbol{n}}$

$$\text{anomalyScore}(I_k) = |\, I_n \,| * \text{matchcost}(I_k, S)$$

instances of $\boldsymbol{I_k}$

- Example:



Experiments (Cargo shipments)

- Data: obtained from Customs and Borders Protection (CBP)
- Scenario:
  - Marijuana seized at Florida port [press release by U.S. Customs Service, 2000].
  - Smuggler did not disclose some financial information, and ship traversed extra port.
  - GBAD-P discovers the extra traversed port;
  - GBAD-MPS discovers the missing financial info.

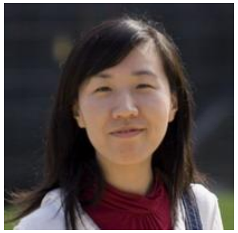
Experiments (Network intrusion)

- Data: 1999 KDD Cup Network Intrusion
  - 100% of attacks were discovered with GBAD-MDL
  - 55.8% for GBAD-P and 47.8% for GBAD-MPS

  Note
    - Data consists of TCP packets that have fixed size
    - Thus, the inclusion of additional structure, or the removal of structure, is not relevant here.
    - Modification is the only relevant one, at which GBAD-MDL performs well

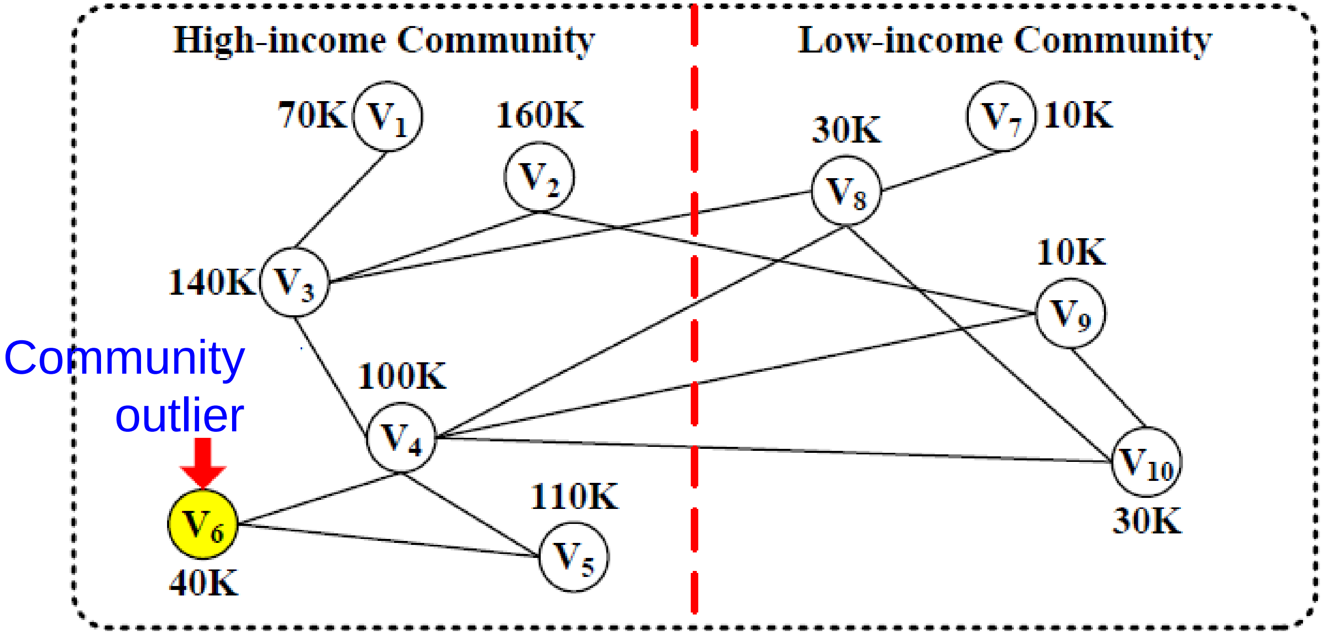
  - High (unreported) false positive rate!

Gao et al. '10



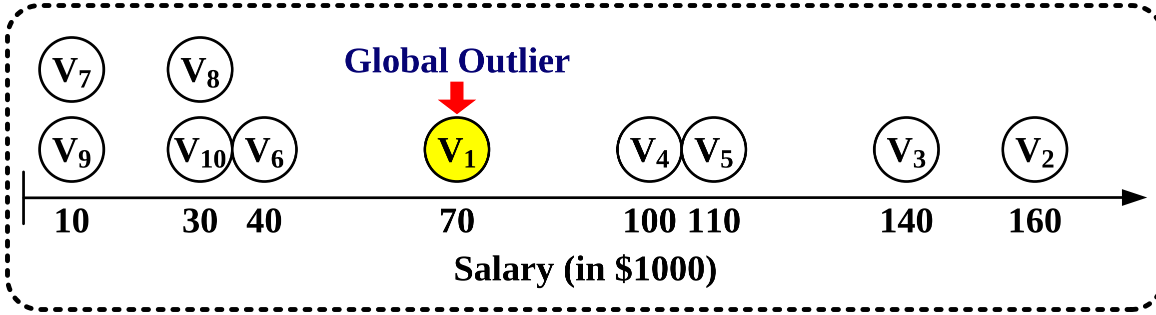
Community Outliers

- Definition
  - Two information sources: links, node features
  - Communities based on **both** links and node features
  - Objects with features deviating from other community members defined as community outliers



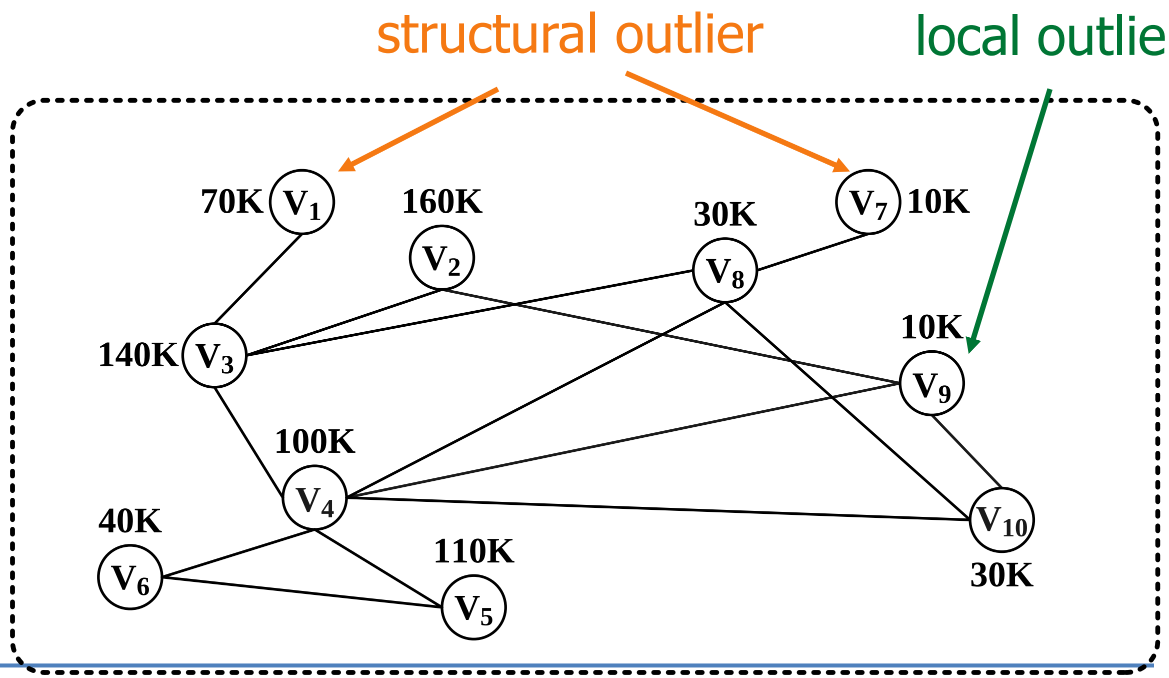
Other network outliers

1) Global outlier:
only considers node features



2) Structural outlier:
only consider links

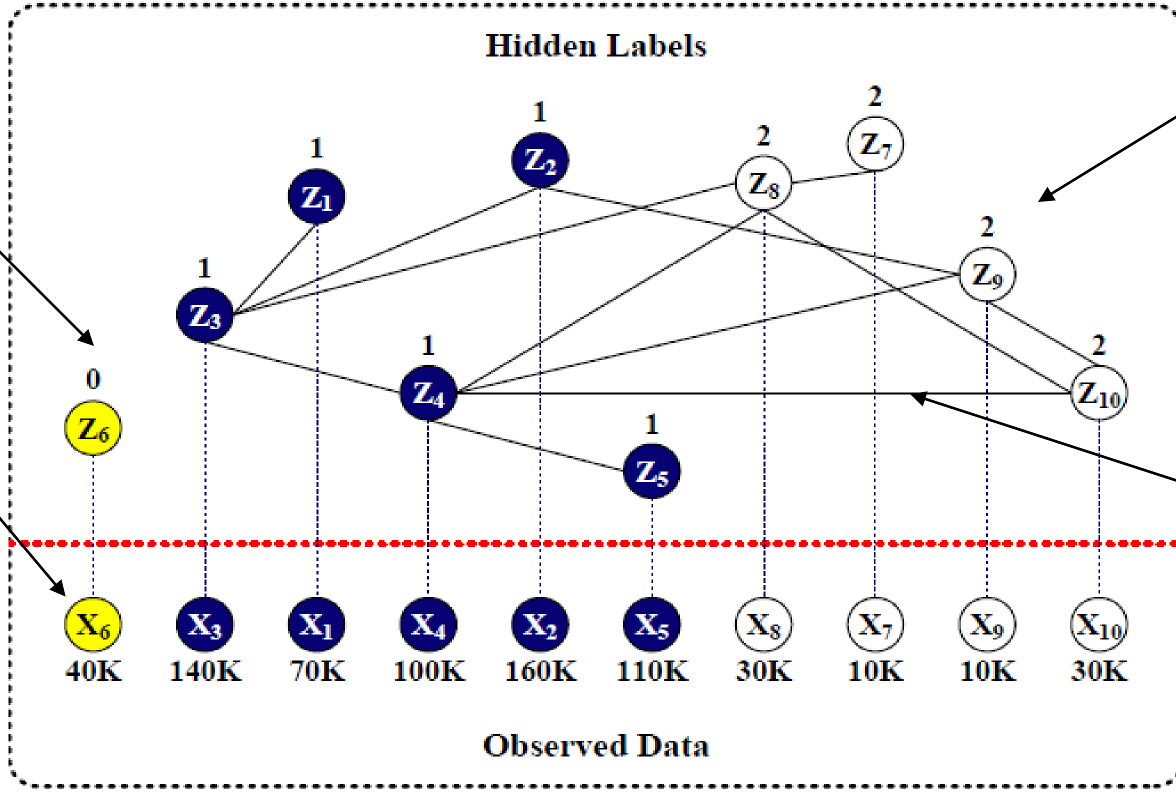
3) Local outlier:
only consider the feature values of direct neighbors



A unified probabilistic model

outlier

node features X



community label Z {0, 1, …, K}

link structure W

$$\Theta = (\theta_1, \ldots, \theta_K)$$

K: number of communities (user input)

model parameters X's are drawn from

Optimization formulation

details

- Maximize P(X) $\propto$ P(X|Z) P(Z)
  - P(X|Z) depends on community label and model param.s
    - e.g., salaries in the high or low-income communities follow Gaussian distributions defined by mean and std

$$P(x_i = s_i | z_i = k) = P(x_i = s_i | \theta_k)$$

Normal with $\{\mu_k, \sigma_k^2\}$

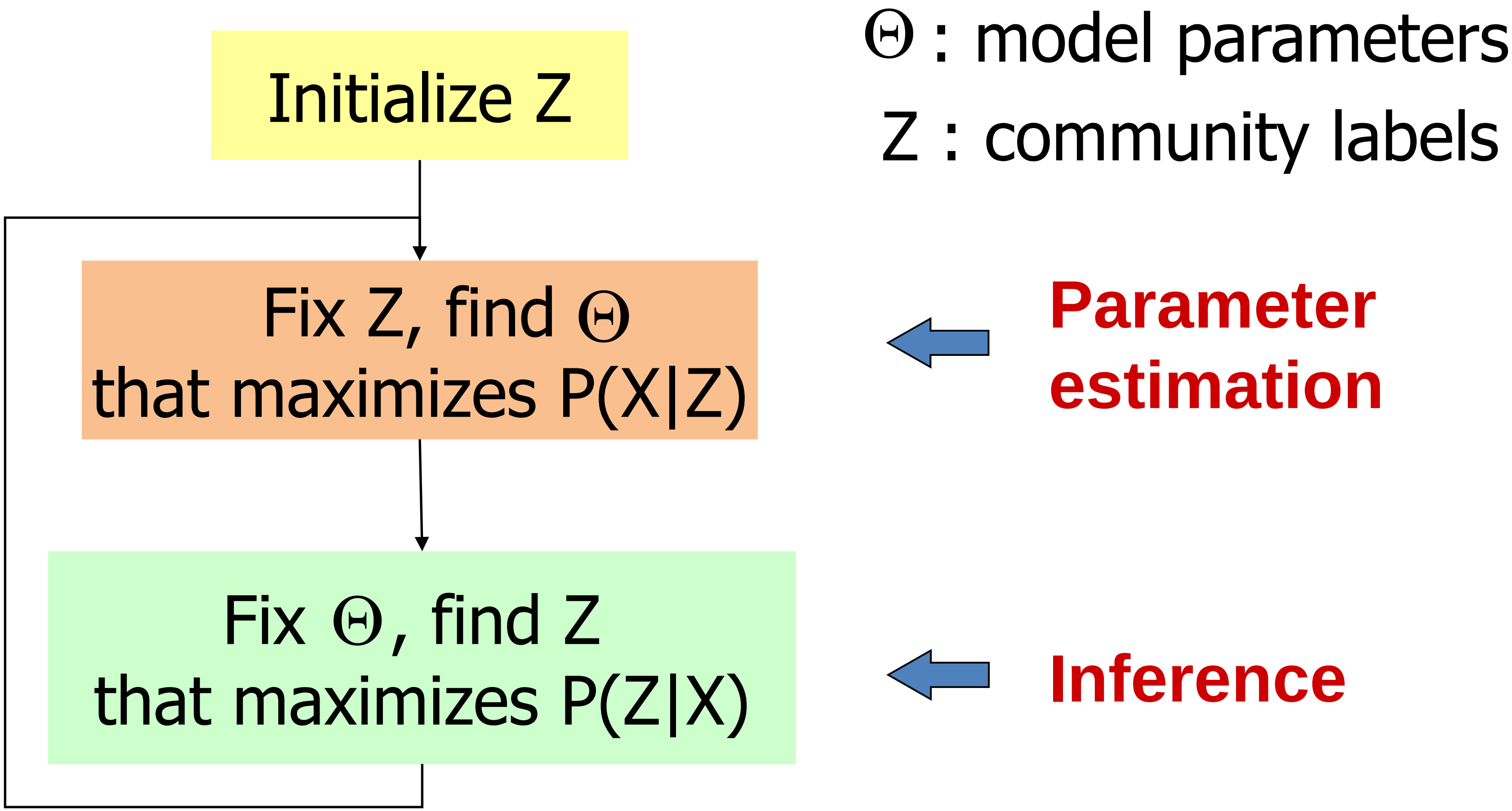
$$P(x_i = s_i | z_i = 0) = \rho_0$$

Uniform for outliers

  - P(Z) is higher if neighboring nodes from normal communities share the same community label
    - e.g., two linked nodes are likely to be in the same community
    - outliers are isolated—does not depend on the labels of neighbors

$$P(Z) \propto \sum_{w_{ij}>0, z_i \neq 0, z_j \neq 0} w_{ij} \delta(z_i - z_j)$$

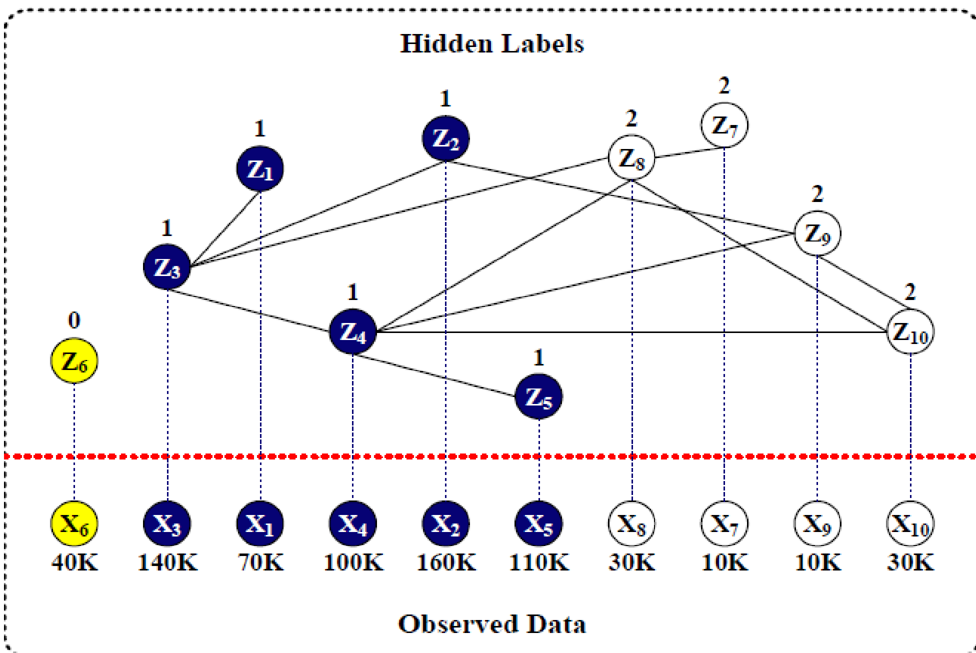
Algorithm



- Initialization is very important (by clustering)
- Convergence/correctness not guaranteed

Algorithm: parameter estimation

- Calculate model parameters Θ
  - maximum likelihood estimation
- Continuous: $\{\mu_k, \sigma_k^2\}$
  - mean: sample mean of the community
  - std: square root of sample variance of community



high-income:
mean: 116k
std: 35k

low-income:
mean: 20k
std: 12k

Algorithm: inference

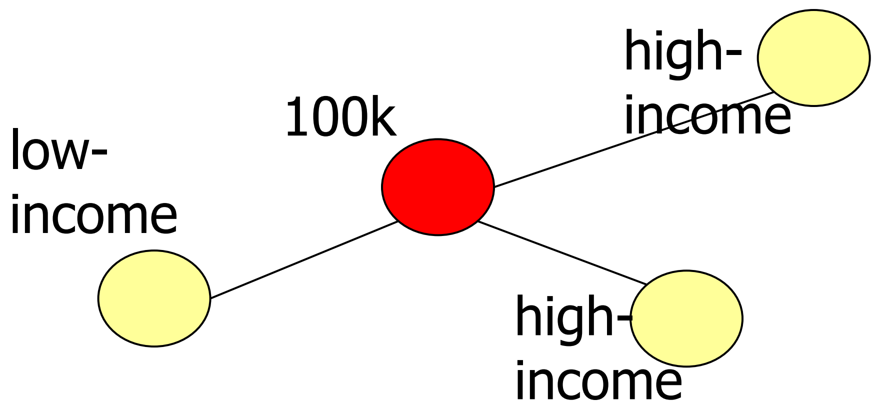
- Calculate label assignments **Z**
  - Model parameters are known
  - Iteratively update the community labels of nodes
  - For each node: select label that maximizes:

$$P(z_i | x_i = s_i, z_{I-\{i\}}) \propto P(x_i = s_i | z_i) \cdot \exp\left(\lambda \sum_{j \in N_i} w_{ij} \delta(z_i - z_j)\right)$$

high-income:	P(salary=100k|high-income)	P(high-income|neighbors)
low-income:	P(salary=100k|low-income)	P(low-income|neighbors)
outlier:	constant	

high-income:
mean: 116k
std: 35k

low-income:
mean: 20k
std: 12k

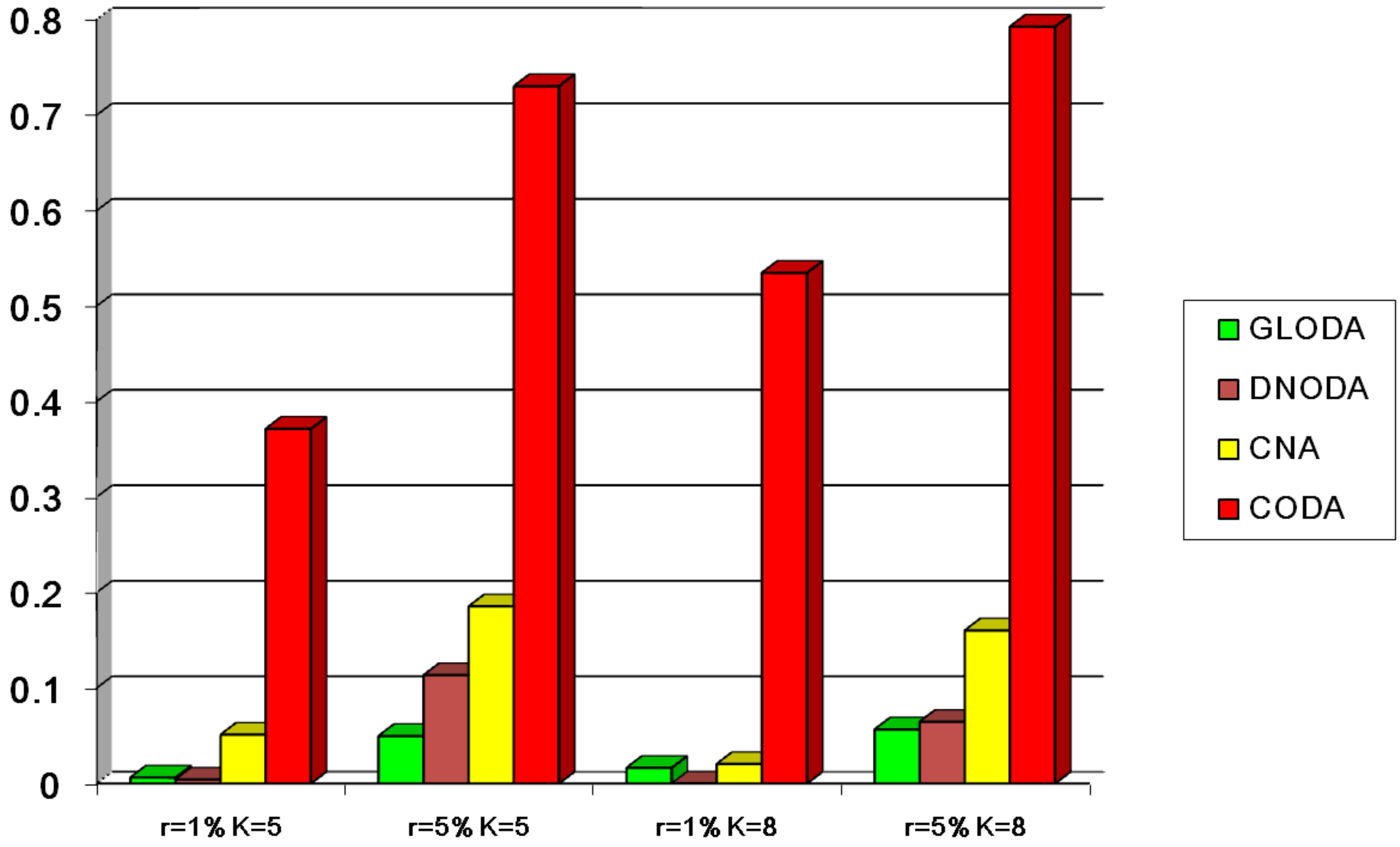


Experiments: Simulations

- Data
  - Generate continuous data based on Gaussian distributions and generate labels according to the model
  - r: percentage of outliers, K: number of communities
- Baseline models
  - GLODA: global outlier detection (based on node features only)
  - DNODA: local outlier detection (check the feature values of direct neighbors)
  - CNA: partition data into communities based on links and then conduct outlier detection in each community

Experiments: Simulations

Precision



Case study on DBLP

- Conferences graph
  - Links: % common authors among two
  - Node features: publication titles in the conference
- Communities:
  - **Database:** ICDE, VLDB, SIGMOD, PODS, EDBT
  - **Artificial Intelligence:** IJCAI, AAAI, ICML, ECML
  - **Data Mining:** KDD, PAKDD, ICDM, PKDD, SDM
  - **Information Analysis:** SIGIR, WWW, ECIR, WSDM
- Community outliers: CVPR and CIKM

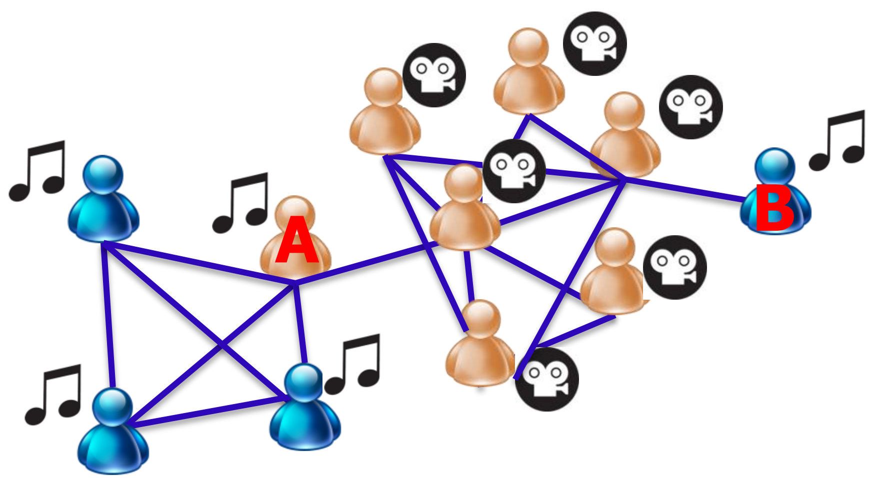
Cohesive groups in attributed graphs

- Problem:

Given a graph with node attributes (features)

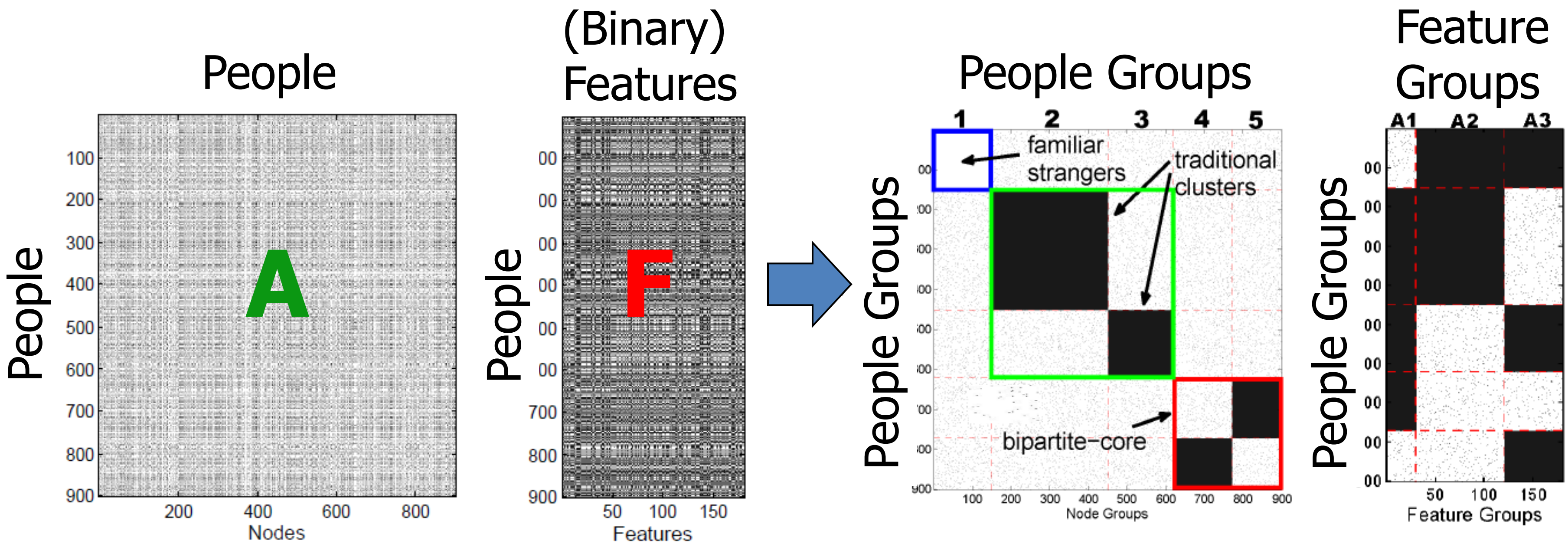
- social networks + user interests
- phone call networks + customer demographics
- gene interaction networks + gene expression info

Find cohesive clusters, bridges, anomalies



Note: cohesive cluster: similar connectivity & attributes

Problem sketch



Given adjacency matrix **A** and feature matrix **F**

Find homogeneous blocks (clusters) in **A** and **F**

* parameter-free
* scalable

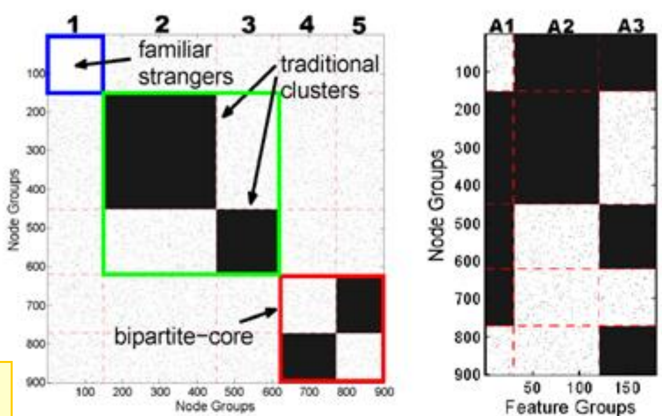
Problem formulation

1. How many node- & attribute-clusters?
2. How to assign nodes and attributes to clusters?

Main idea: employ Minimum Description Length

$$\underbrace{L\,(M)}_{\text{encoding length of clustering}} + \underbrace{L\,(D|M)}_{\text{encoding length of blocks}}$$

Good Clustering ⟺ implies ⟺ Good Compression



Problem formulation

details

- L (M) : Model description cost
  1. $\log^\star n + \log^\star f$ n: #nodes, f: #attributes
  2. $\log^\star k + \log^\star l$ k: #node-clusters, l: #attribute-clusters
  3. $nH(P) + fH(Q)$

$$p_i = \frac{r_i}{n}$$ ← size of node-cluster i

$$q_j = \frac{c_j}{f}$$ ← size of attribute-cluster j

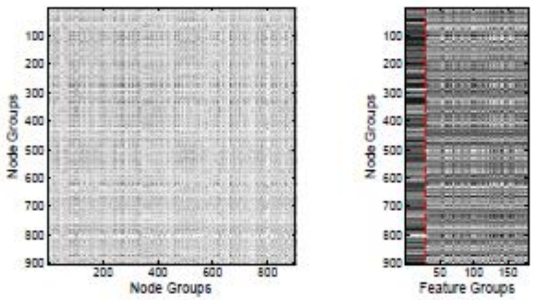
- L(D|M): Data description cost given Model
  1. For each block in A and F , #1s: $\log^\star n_1(B_{ij})$
  2. $E(B_{ij}) = -n_1(B_{ij}) \log_2(P_{ij}(1)) - n_0(B_{ij}) \log_2(P_{ij}(0))$

where $P_{ij}(1) = n_1(B_{ij})/n(B_{ij})$

$r_i c_j$ or $r_i r_j$

A similar problem (column re-ordering for minimum total run length) is shown to be NP-hard [Johnson+].
(reduction from Hamiltonian Path)

Algorithm sketch

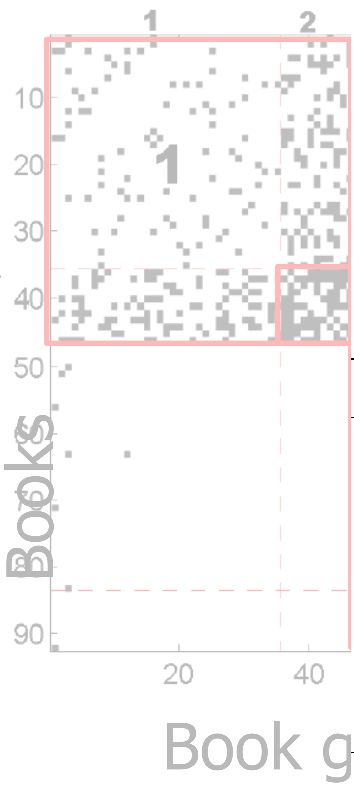


(a) k=1 l=2

`Split-FeatureGroup`

The algorithm is iterative and monotonic –will converge to local optimum

PICS at work (Political books)



Examples of "core" liberal and conservative books

Liberal
–Lies and the Lying Liars Who Tell Them: *A Fair and Balanced Look at the Right*
–Big Lies: *The Right-Wing Propaganda Machine and How It Distorts the Truth*
–The Lies of George W. Bush
–Dude, Where's My Country?

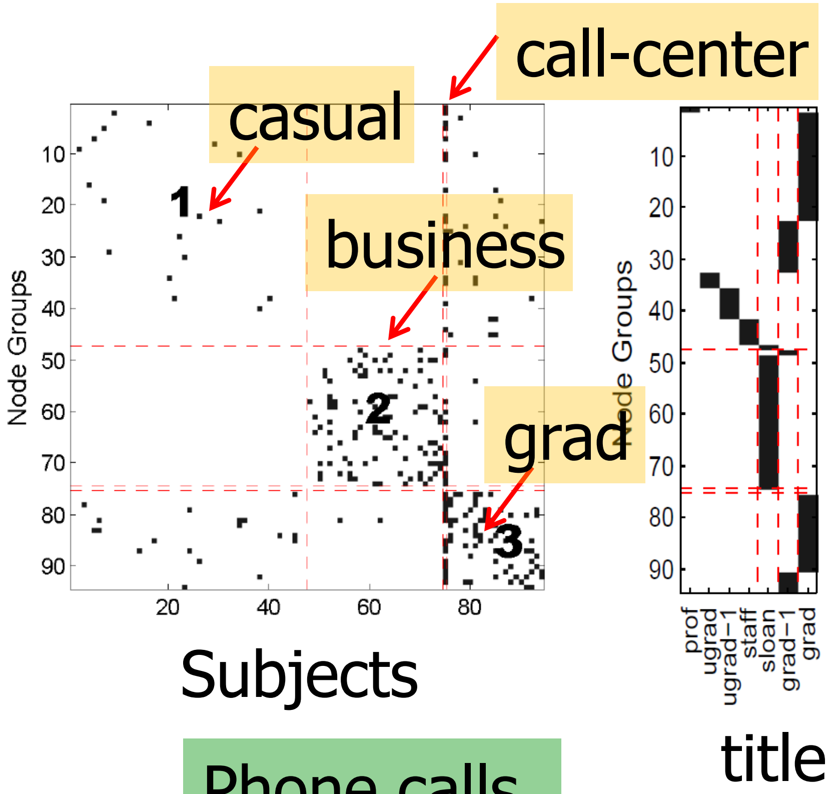
Conservative
–Persecution: *How Liberals Are Waging War Against Christianity*
–Deliver Us from Evil: *Defeating Terrorism, Despotism, and Liberalism*
–Tales from the Left Coast
–A National Party No More

Examples of bridging 'conservative' books

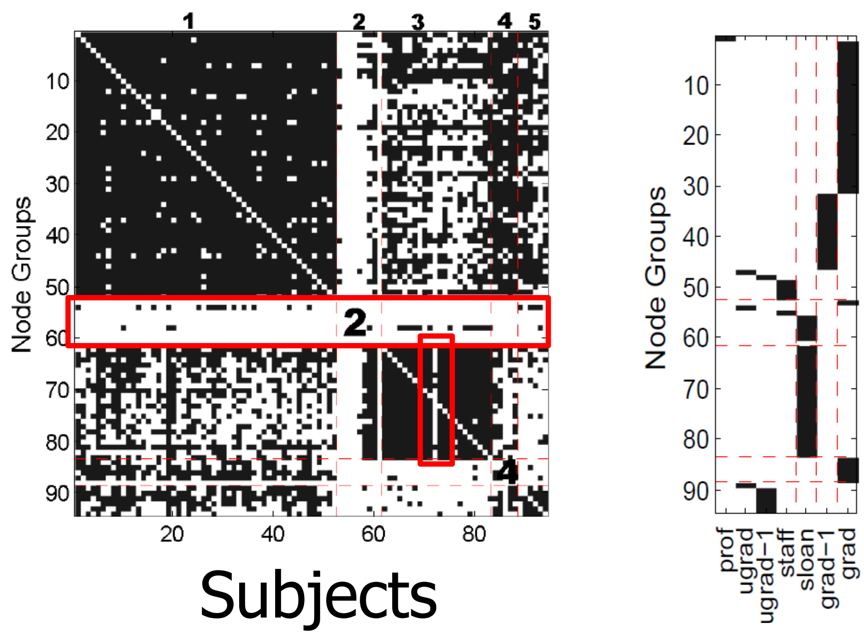
–*Bush at War*
–*The Bushes: Portrait of a Dynasty*
–*Rise of the Vulcans: The History of Bush's War Cabinet*

"core and perip

PICS at work (Reality mining)

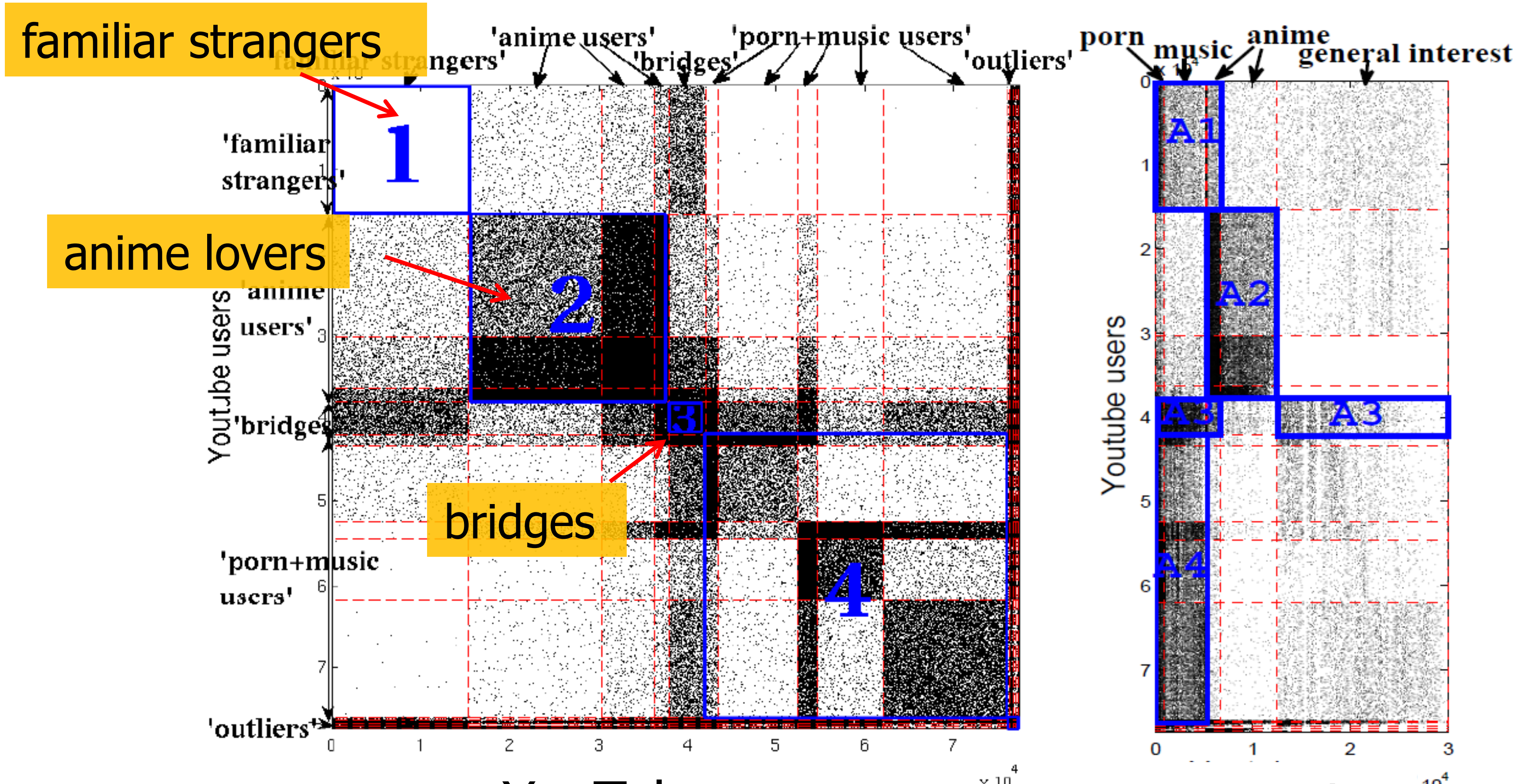


Phone calls



Device scans

PICS at work (YouTube)



YouTube users

YouTube groups

77K users
30K groups

Part I: References (attribute graphs)

Substructures

Community mining

- C. C. Noble and D. J. Cook. Graph-based anomaly detection. KDD, pages 631–636, 2003.
- W. Eberle and L. B. Holder. Discovering structural anomalies in graph-based data. ICDM Workshops, pages 393–398, 2007.
- Michael Davis, Weiru Liu, Paul Miller, George Redpath: Detecting anomalies in graphs with numeric labels. 1197-1202, CIKM 2011.
- Jing Gao, Feng Liang, Wei Fan, Chi Wang, Yizhou Sun, Jiawei Han: On community outliers and their efficient detection in information networks. KDD 2010: 813-822.
- Leman Akoglu, Hanghang Tong, Brendan Meeder, Christos Faloutsos. PICS: Parameter-free Identification of Cohesive Subgroups in large attributed graphs. SDM, 2012.

Tutorial Outline

- Motivation, applications, challenges
- **Part I:** Anomaly detection in **static** data
  - Overview: Outliers in **clouds of points**
  - Anomaly detection in **graph data**
- **Part II:** Event detection in **dynamic** data
  - Overview: Change detection in **time series**
  - Event detection in **graph sequences**
- **Part III:** Graph-based **algorithms and apps**
  - Algorithms: **relational learning**
  - Applications: **fraud and spam** detection


Out to
LUNCH!