An Embedding Approach to Anomaly Detection

Renjun Hu[1], Charu Aggarwal[2], Shuai Ma[1], and Jinpeng Huai[1]

[1]SKLSDE Lab, Beihang University, China

[2]IBM T. J. Watson Research Center, USA

Motivation

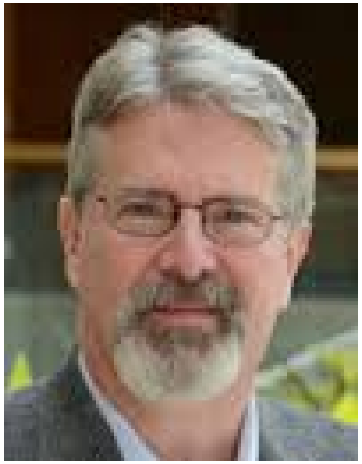
- Anomaly detection
  - Identification of patterns in data that do not conform to expected behaviors [Chandola et al. 2009]
  - Useful in a wide variety of applications

- In networks, anomaly detection has broader meanings
  - Application-specific significance
  - Possibility to improve the performance of network-centric mining tasks such as community detection and classification

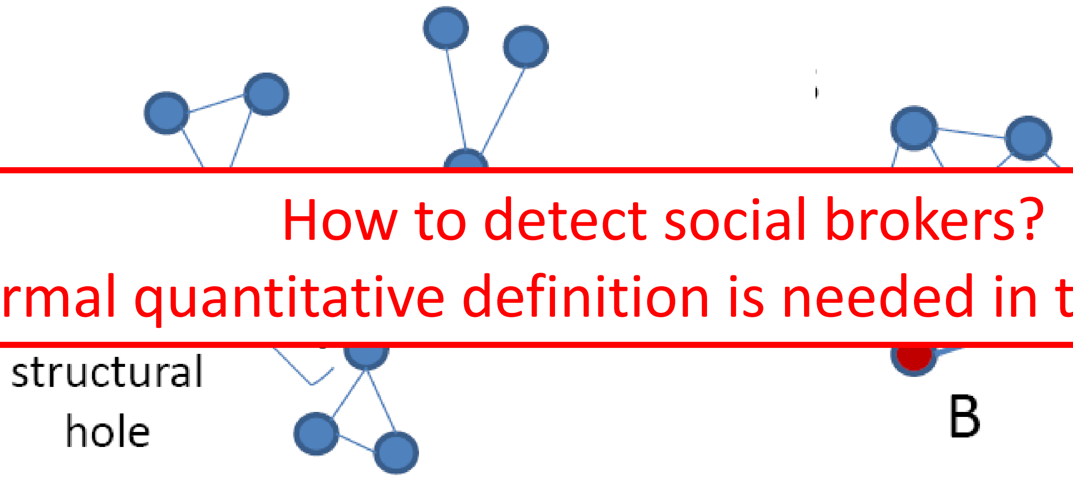
V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Comput. Surv.* 41(3), 2009.

Motivation

- Structural hole theory [Burt 1992, 2004]
  - Theory of social capital
  - A structural hole is a gap between two nodes who have complementary sources to information



Prof. Ronald S. Burt



How to detect social brokers?

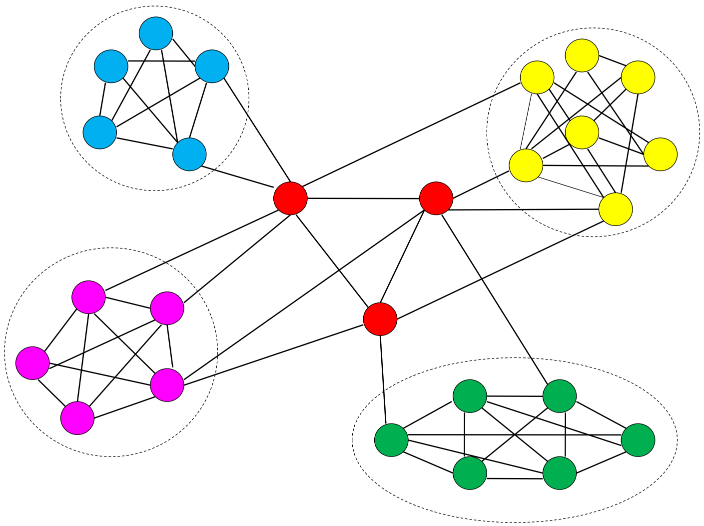
A formal quantitative definition is needed in the first place!

- Node A (social broker) is more likely to get novel information than B, even though they have the same number of links.

Burt, Ronald S. (1992). Structural holes: the social structure of competition. Harvard University Press.
Burt, Ronald S. (2004). Structural Holes and Good Ideas. *American Journal of Sociology* 110 (2): 349–399.

Motivation

- Structural inconsistencies
  - Nodes that connect to a number of diverse influential communities
  - Detect social brokers quantitatively

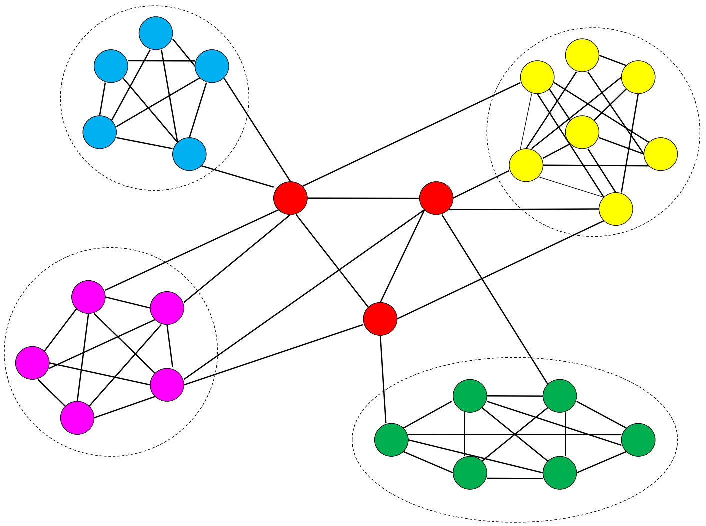


- Anomalousness from homophily [McPherson et al. 2001]
  - Linked nodes have similar properties
  - Fundamental to a wide variety of algorithms in network science
    - ✓ *E.g.*, community detection, collective classification, link prediction, influence analysis
  - Violated by structural inconsistencies

M. McPherson, L. Simth-lovin and J. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, Vol. 27: 415-444, 2001.

Motivation

- Structural inconsistencies
  - Nodes that connect to a number of diverse influential communities
  - Detect social brokers quantitatively



- The presence of structural inconsistencies may:
  - have a substantial impact on network structure
    - ✓ *E.g.*, all nodes tend to form one large cluster
  - prevent effective applications of network mining algorithms
    - ✓ *E.g.*, hard for community detection algorithms to achieve meaningful clusters

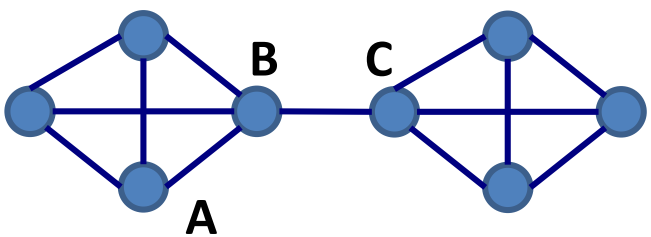
Outline

- Anomaly detection model
  - Graph embedding
  - A quantitative measure of anomaly
- Algorithm optimization techniques
- Evaluation

Why graph embedding?

- Structural inconsistencies
  - connect to a number of diverse influential communities

- Evaluate the diversity or similarity of nodes. How?



- To node B, node A is more similar than C, even though they have the same (global) distance from B.

- Graph embedding
  - Associate each node with a multidimensional vector
  - Preserve local linkage structure (instead of global structure)
  - Each dimension corresponds to a community in the network

Why graph embedding?

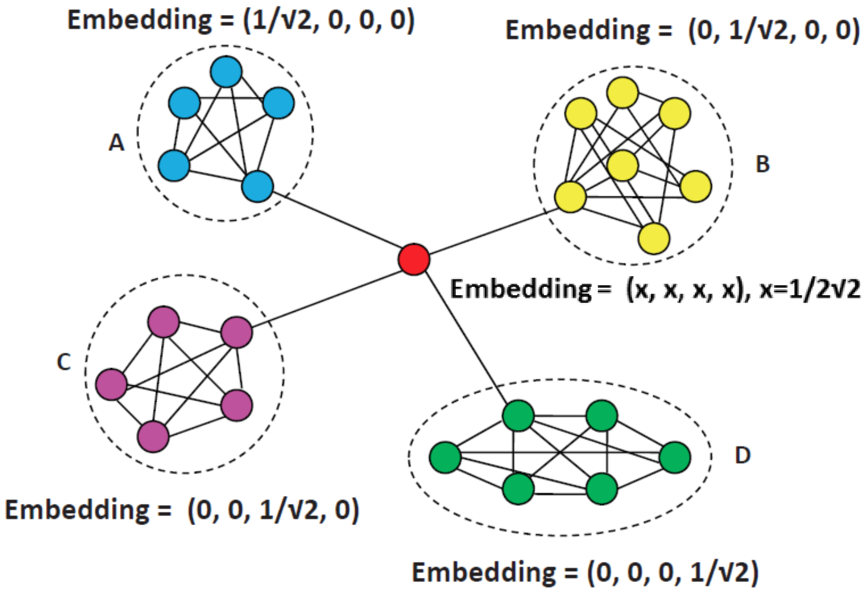
- Structural inconsistencies
  - connect to a number of diverse influential communities

- An alternative option: doing community detection followed by anomaly detection
  - Do not distinguish anomalies from normal nodes
  - The presence of anomalies has certain impacts on the results of community detection
  - Community detection is a heavy task.
  - Fail to detect structural inconsistencies!

Graph embedding

- Given an undirected graph G=(V, E), associate each node i with a d-dimensional vector X_i

  - V = {1,2,…,n}
  - d : number of communities
  - X_i : correlation between node i and the d communities



A reasonable selection of d suffices for anomaly detection.
Not necessary to use the number of real-life communities.

Graph embedding

- Given an undirected graph $G=(V, E)$, associate each node i with a d-dimensional vector X_i

- Goal: preserve local linkage structure
  - Connected nodes should have similar values of X_i
  - Disconnected nodes should have diverse values of X_i

- Computation: minimizing objective function O

$$O = \sum_{(i,j)\in E} \left\|X_i - X_j\right\|^2 + \alpha \cdot \sum_{(i,j)\notin E} \left(1 - \left\|X_i - X_j\right\|\right)^2, \alpha = \frac{m}{\binom{n}{2} - m}$$

  - n: number of nodes in G, m: number of edges in G
  - α : balancing factor that regulates the importance of the two components in O
  - The embedding ensures that $0 \leq \|X_i - X_j\|^2 \leq 1$

A quantitative measure

- Inspired by structural inconsistencies and structural holes (social brokers)
  - Connect to a number of diverse influential communities
  - Bridge across complementary sources

- *NB(i)*: how node *i* connects to communities

$$NB(i) = \left(y_i^1, \ldots, y_i^d\right) = \sum_{(i,j)\in E} \left(1 - \left\|X_i - X_j\right\|\right) \cdot X_j$$

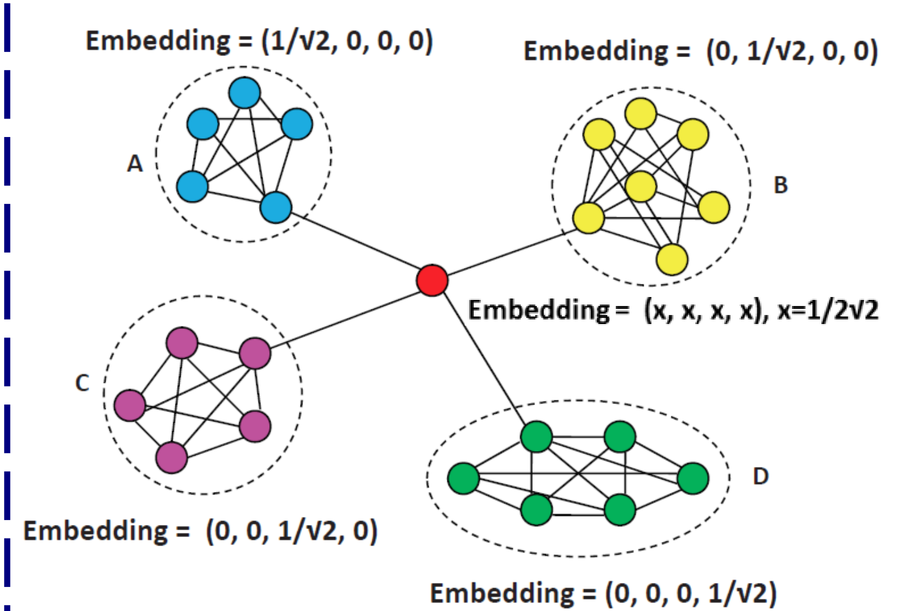
- *AScore(i)*: the anomalousness of node *i*

$$AScore(i) = \sum_{k=1}^{d} \frac{y_i^k}{y_i^*}, \; y_i^* = \max\left\{y_i^1, \ldots, y_i^d\right\}$$

  - Detect anomalies by *AScore*(*i*) > *thre*

Example

- Optimality of embedding, *i.e.*, minimum value of O
  - Small values within groups because of missing edges
  - No values across groups
  - Certain values for the red node (no better embedding)



- Anomalousness of nodes
  - *AScore(red)* = 4 (equal values in dimensions of *NB*(*red*))
  - *AScore(i)* ≈ 1 for others (*NB*(*i*) only has a dominating dimension)

$$O = \sum_{(i,j)\in E} \|X_i - X_j\|^2 + \alpha \cdot \sum_{(i,j)\notin E} \left(1 - \|X_i - X_j\|\right)^2$$

$$AScore(i) = \sum_{k=1}^{d} \frac{y_i^k}{y_i^*},\ y_i^* = \max\left\{y_i^1, \ldots, y_i^d\right\}$$

The red node is detected as an anomaly!

Outline

- Anomaly detection model

- Algorithm optimization techniques
  - Sampling
  - Graph partitioning based initialization
  - Dimension reduction

- Evaluation

Issues in the model

- Objective function O is a sum over $O(n^2)$ terms
  - Forbidden in large social networks
- Optimizing O uses a gradient descent method
  - Critically dependent on a good initialization
- Dimensionality of embedding (*i.e.*, d) could be large
  - *E.g.*, 8,353 for YouTube and 6,288,363 for Orkut [Yang & Leskovec 2012]

J. Yang and J. Leskovec. Defining and evaluation network communities based on ground-truth. In *ICDM*, 2012.

Sampling

- Objective function O is a sum over $O(n^2)$ terms

$$O = \sum_{(i,j)\in E} \|X_i - X_j\|^2 + \alpha \cdot \sum_{(i,j)\notin E} \left(1 - \|X_i - X_j\|\right)^2, \alpha = \frac{m}{\binom{n}{2} - m}$$

- Observation: balancing factor α is close to 0
  - Very inefficient
  - Possible to approximately represent O by sampling

- Sampled objective function O

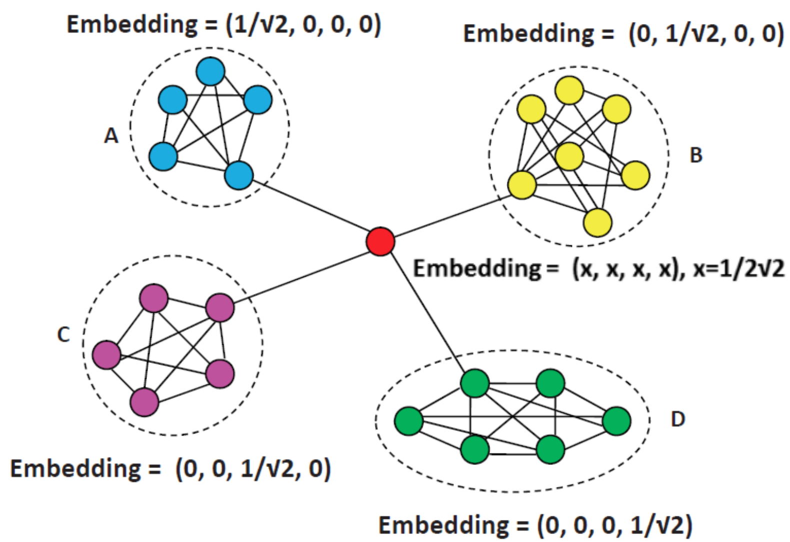
$$O \approx \sum_{(i,j)\in E} \|X_i - X_j\|^2 + \sum_{(i,j)\in E_s} \left(1 - \|X_i - X_j\|\right)^2, E_s \subset \{(i,j) \mid (i,j) \notin E\}$$

  - $|E_s| = |E| = m$

Graph partitioning based initialization

- Optimizing O uses a gradient descent method
  - Critically dependent on a good initialization

- A good initialization means small value of O
  - Densely connected nodes have similar values of X_i
  - Nodes across groups have diverse values of X_i



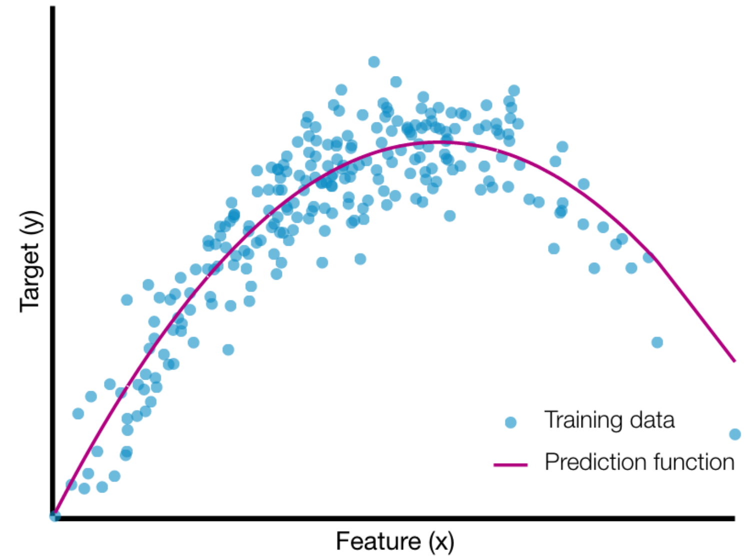
- Incorporating graph partitioning (METIS) for initialization
  - P_i : partition number of node i

$$X_i = (x_i^1, \ldots, x_i^d), x_i^j = \begin{cases} 1/\sqrt{2} & j = P_i \\ 0 & j \neq P_i \end{cases}$$

Dimension reduction

- Dimensionality of embedding (*i.e.*, *d*) can be large

- The complete d-dimensions are unnecessary
  - Nodes typically connect to a limited number of communities
  - A limited number of communities suffice to ascertain anomalies



(Gordon) Hughes Effect

- Data approximation (k+β reduction)
  - only maintain (k+β)-dimensions for embedding of each node
  - k : the maximum number of communities to connect
  - β : tolerate mistakes when determining the k communities
  - $k \ll d$ & $\beta \ll d$, *e.g.*, 10 & 2 for a network with $n = 10^6$

Impacts of optimization techniques

	Space	Efficiency	Effectiveness
Sampling	/	Prev.: O($n^2 \cdot d$) After: O($m \cdot d$)	Remain effective (from experiments)
Graph partitioning	/	Prev.: 0 After: O($n+m+d \cdot \log(d)$)	Provide a good initialization
k+β reduction	Prev.: O($n \cdot d$) After: O($n \cdot (k+\beta)$)	Prev.: O($t \cdot m \cdot d$) t : # of iterations After: O($t \cdot m \cdot (k+\beta)$)	Slightly improve effectiveness

Outline

- Anomaly detection model
- Algorithm optimizations
- Evaluation

Experimental settings

- Datasets

Dataset	# of nodes	# of edges	Descriptions
Amazon	334,863	925,872	Product co-purchasing
DBLP	1,150,852	5,098,175	Co-authorship
Synthetic	10^5 - 4×10^6	$m = n^{1.15}$	LFR-benchmark graph

  - Anomaly injection on Synthetic data for ground-truth of anomalies

- Algorithms
  - Embed(d) : embedding of d-dimensions
  - Embed(k+β) : embedding with k+β reduction
  - Oddball : based on violation of power-laws of egonet-based features
  - MDS(d) : similar to Embed(d), except using multi-dimensional scaling for embedding (preserve global structure)

- Parameters: $d = n/500$, $k = avgDeg$, $\beta = k/4$

- Implementation: C++, Core i5 3.10GHz, 16GB of memory

Case study on DBLP

- Different people with the same name

  Wei Wang

  - 84 people named Wei Wang [DBLP, May 10 2016]
  - University of Waterloo (Canada), Fudan University (China), University of California, San Diego (USA), etc.

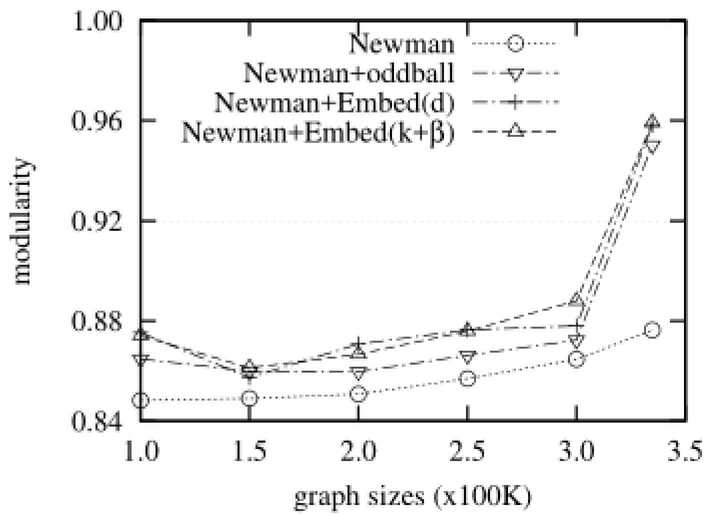
- People with many collaborators in diverse institutes

  Dr. Ajith Abraham

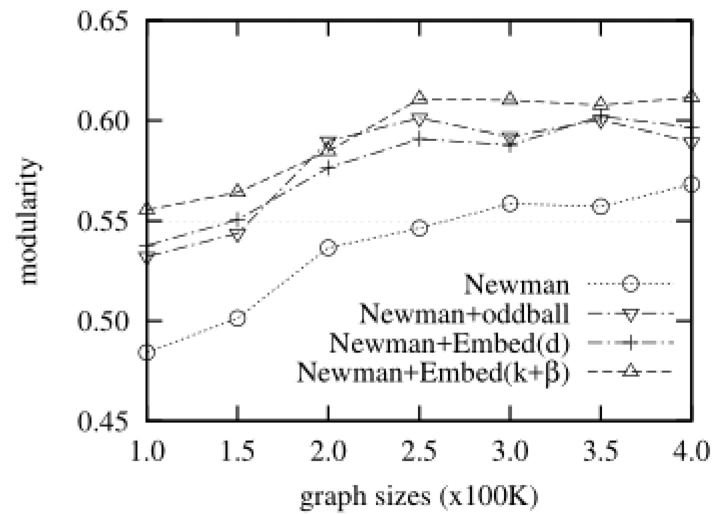
  - Director of intelligence research labs which has members from more than 100 countries
  - Work in a multi-disciplinary environment involving machine intelligence, cyber security, sensor networks and data mining
  - Teach in 23 universities all over the world

Quality study: modularity

- Modularity measures the strength of division of a network into communities
- Using modularity to evaluate the improvement of the effectiveness of community detection



(a) AMAZON dataset



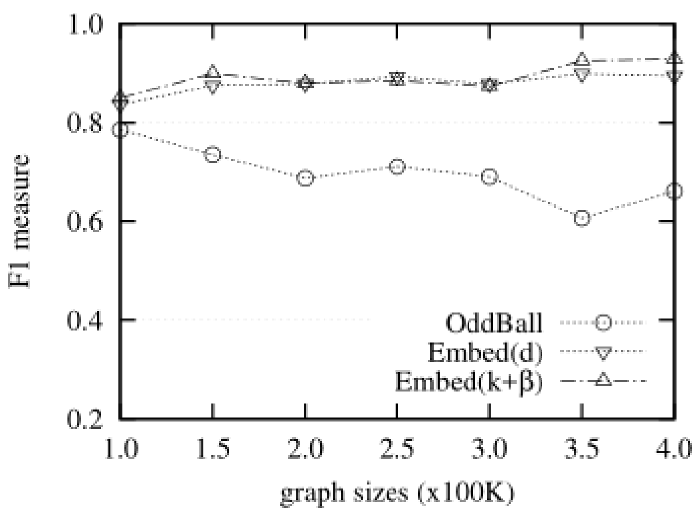
(b) DBLP dataset

	oddball	**Embed(*d*)**	**Embed(*k*+β)**
Amazon	2.1%	2.8%	**3.0%**
DBLP	4.2%	4.1%	**5.6%**

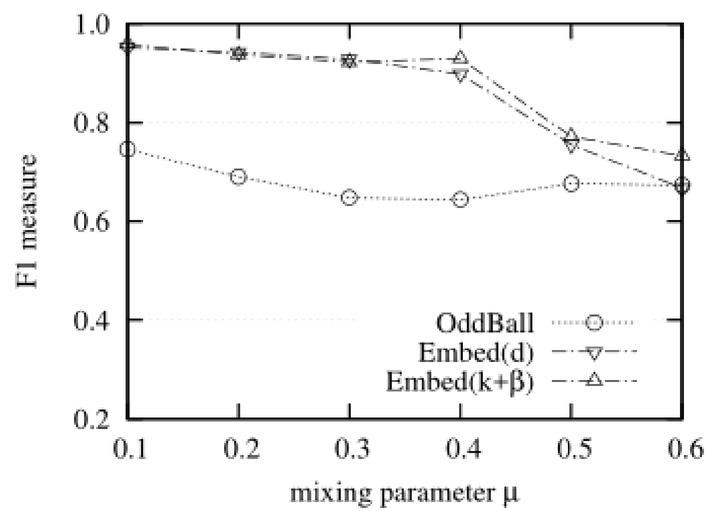
Table 1: Improvement of modularity

Quality study: F_1 measure

- On Synthetic data with ground-truth of anomalies
- Mixing parameter μ: fraction of inter-group edges (*i.e.*, μ ↑, strength of community structure ↓)



(a) $k + \beta$ reduction



(b) The mixing parameter μ

	oddball	Embed(*d*)	Embed(*k*+β)
Varying graph sizes	70%	88%	**89%**
Varying μ	68%	86%	**88%**

Table 2: F_1 score of anomalies

Impacts on quality: *d* & embedding

- Synthetic data, n = 400K, n/500 = 800

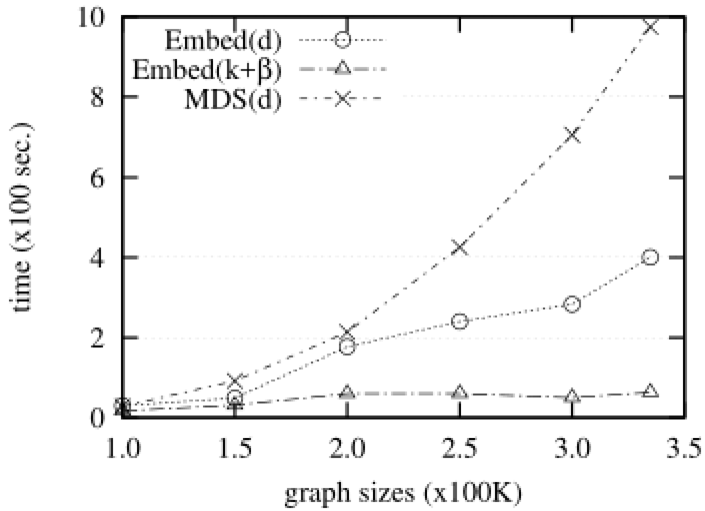
	MDS(*d*)	**Embed(*d*)**
d = 200	11.3%	**89.4%**
d = 400	13.6%	**90.6%**
d = 600	12.7%	**89.8%**
d = 800	7.9%	**85.5%**
d = 1000	11.3%	**88.8%**
Average	11.3%	**88.8%**

Table 3: MDS(*d*) vs. Embed(*d*) using F_1 measure

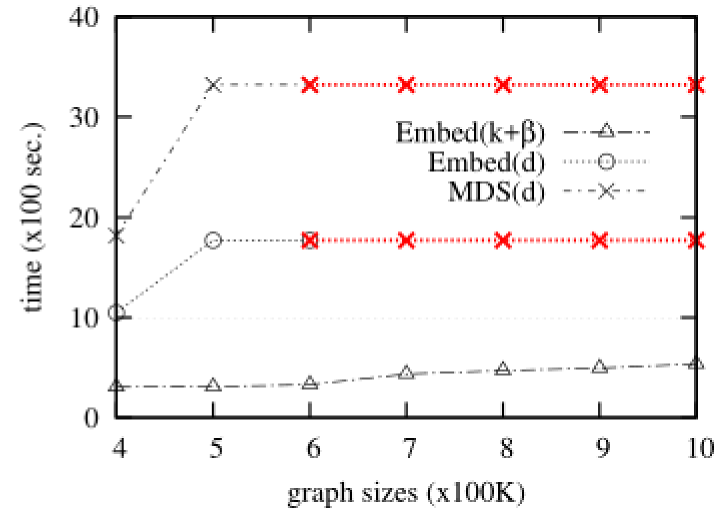
- Multi-dimensional scaling fails to effectively detect anomalies
- Our approach works well as long as *d* falls into a reasonable range

Efficiency study

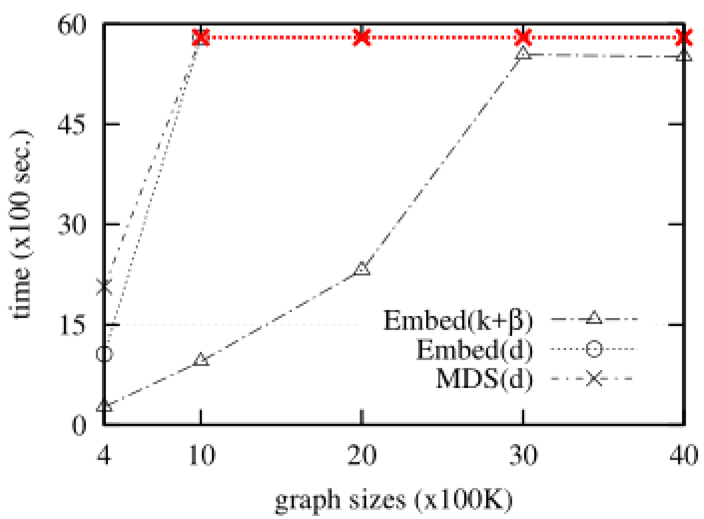
x : out of memory exception



(a) AMAZON dataset



(b) DBLP dataset



(c) SYNTHETIC dataset

	E(k+β)/E(d)	E(k+β)/MDS(d)
Amazon	35.3%	25.0%
DBLP	23.4%	13.1%
Synthetic	25.6%	13.2%

Table 4: running time comparison

Summary

- Structural inconsistencies
  - Nodes that connect to a number of diverse influential communities
  - A formal quantitative definition of social brokers
- An embedding approach
  - Preserve local linkage structure of networks
  - A quantitative measure *Ascore* inspired by structural inconsistencies and structural holes
  - Three algorithm optimization techniques
- Quality and efficiency results
  - Modularity increases 2.9%, 4.9% and 6.9% on Amazon, DBLP and Synthetic data
  - F1 measure is 88% on Synthetic data
  - Running time increases reasonably *w.r.t* graph sizes

Thanks!

Q & A