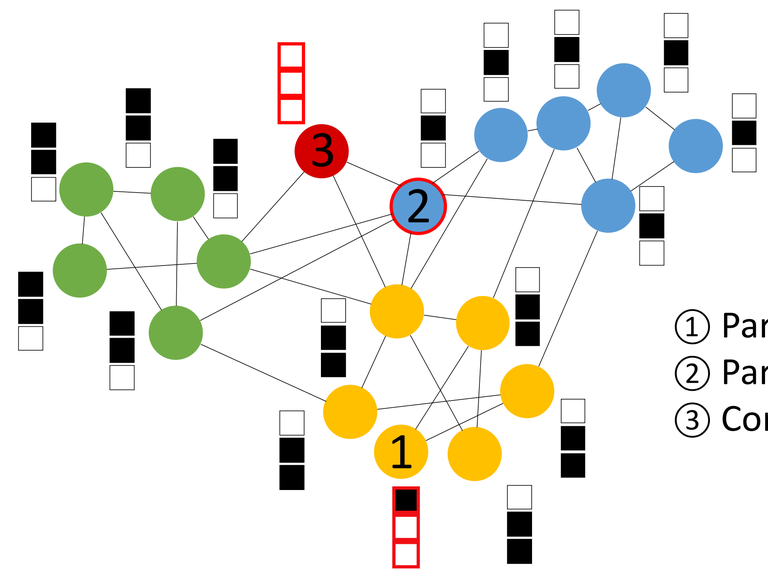
Professorship of Data Mining and Analytics
Department of Informatics
Technical University of Munich

Bayesian Robust Attributed Graph Clustering
Joint Learning of Partial Anomalies and Group Structure

Aleksandar Bojchevski, Stephan Günnemann

Motivation & Core Idea

- Goal: use the graph structure and node attributes to group similar nodes together
- Problem: *anomalies* obfuscate the latent clusters in real-world data
- Previous approaches: detect and discard the anomalous nodes
- Key insight: anomalies can materialize only *partially*
  - For a given node, only one source of information is anomalous
- Solution: derive meaningful clusters based on the clean source
  - Jointly perform clustering and anomaly detection



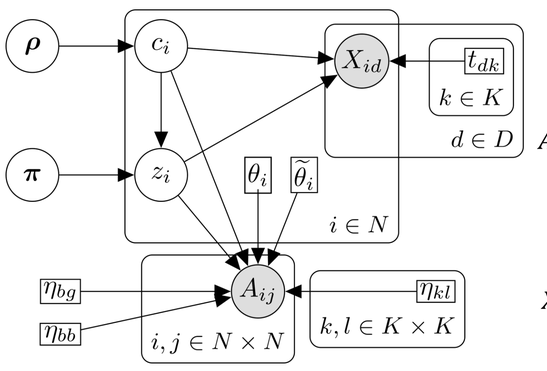
① Partial attribute anomaly
② Partial graph anomaly
③ Complete anomaly

Example: social network user shows anomalous attributes
(to e.g. hide her identity) but her friendship relations are normal

PAICAN: Probabilistic Model

Partial Anomaly Identification and Clustering in Attributed Networks

- Clean graph: Degree-corrected Stochastic Block Model
- Clean binary attributes: Bernoulli Mixture Model



$$A_{ij} \sim \begin{cases} Poisson\left(\theta_i \theta_j \eta_{z_i z_j}\right) & both\ nodes\ clean \\ Poisson\left(\tilde{\theta}_i \eta_{bg}\right) & one\ anomalous \\ Poisson\left(\tilde{\theta}_i \tilde{\theta}_j \eta_{bb}\right) & both\ anomalous \end{cases}$$

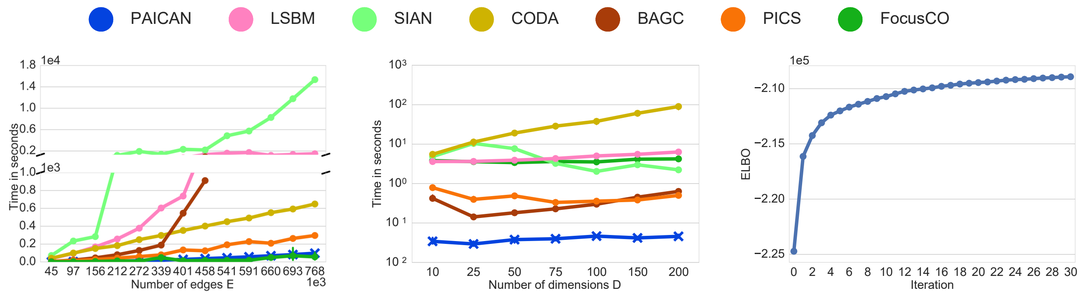
$$X_{id} \sim \begin{cases} Bernouli\left(t_{dz_i}\right) & clean\ \ node \\ Bernouli\ (const.) & anomalous\ node \end{cases}$$

- Separate parameters for edges with a (partially) anomalous node
- Separate parameters for the attributes of anomalous nodes

Algorithmic Solution

Efficient Variational Expectation-Maximization

- We reduce the inference cost from $O(N^2)$ to $O(E)$:
  - Show that certain $O(N)$ terms can be evaluated in $O(1)$
  - Show that in the limit case when the graph grows $(N \to \infty)$ certain terms become negligible with error at most $1/N$
- Linear scaling with the number of edges and attributes

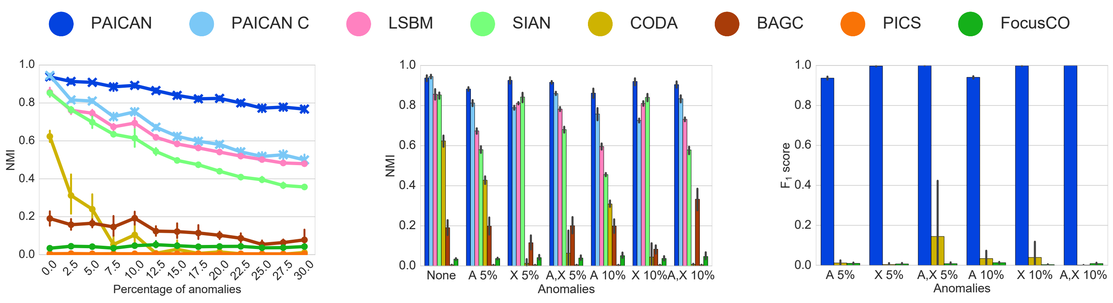


Results

- Clustering performance comparison (NMI) for real-world datasets

	CODA	FocusCO	BAGC	PICS	LSBM	SIAN	PAICAN
Lawyers	0.50	0.28	0.14	0.27	0.50	0.58	**0.66**
Parliament	0.06	0.00	0.53	0.47	0.77	0.73	**0.78**
Cora	d.n.f.	0.13	0.15	0.04	0.52	0.39	**0.53**
Soc. Papers	d.n.f.	0.25	0.17	0.10	0.50	d.n.f.	**0.52**
HVR	0.71	0.50	0.18	0.44	0.83	0.77	**0.89**

- Clustering and anomaly detection performance with increasing percentage of anomalies



- Case study: Anomaly detection

Example partial *attribute* anomaly:
DBLB co-authorship network

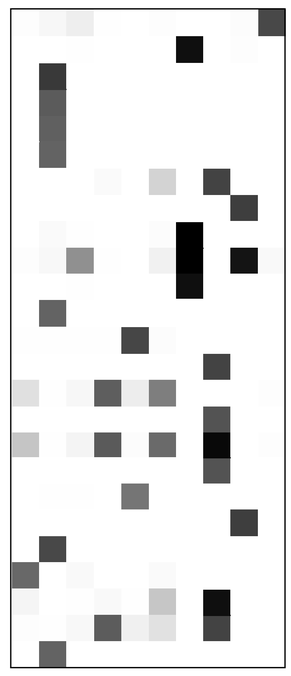
- Srinivasan Parthasarathy published in 18 conferences (e.g. EDBT, IJCAI)
- Most of his co-authors published in just a few (mainly KDD, ICDM, SDM)

Example partial *graph* anomaly:
Amazon co-purchase graph

- The movie “Frozen” belongs to the cluster of movies
- Yet it has multiple edges to nodes from the cluster of children’s clothes

- Case study: Clustering

Discovered topics in the Amazon graph



AAAI-18
February 2 – 7, 2018
New Orleans, Louisiana - USA

www.kdd.in.tum.de/PAICAN