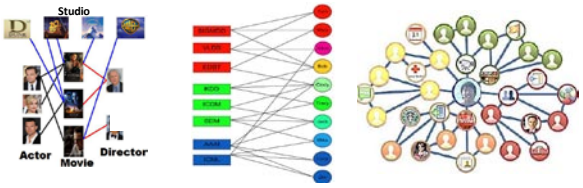
Community Distribution Outliers in Heterogeneous Information Networks

Manish Gupta
Microsoft, gmanish@microsoft.com

Jing Gao
SUNY, Buffalo, jing@buffalo.edu

Jiawei Han
UIUC, hanj@illinois.edu

Heterogeneous Networks are Ubiquitous

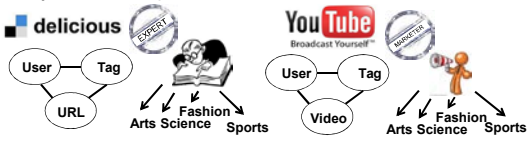


IMDB Network | DBLP Network | Facebook Network

Problem

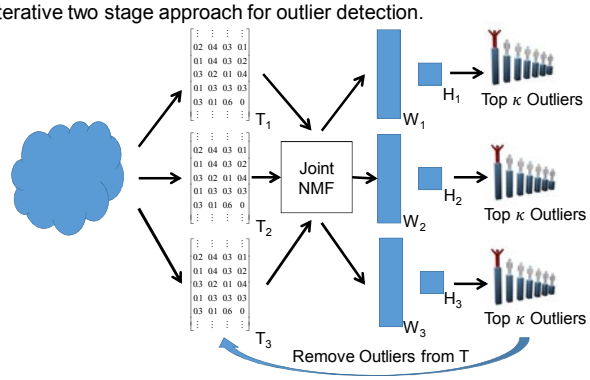
Community Distribution Outliers (CDOutliers) for heterogeneous information networks: Objects whose community distribution does not follow any of the popular community distribution patterns.

Examples



State of the art: Community outliers for homogeneous networks algorithms [Gao et al., 2010], [Gupta et al., 2012a], [Gupta et al., 2012b]. We study community outliers in heterogeneous networks.

Major Idea: We propose a novel joint NMF optimization framework to learn distribution patterns across multiple object types. Further, we propose an iterative two stage approach for outlier detection.



Distribution Pattern for a Type: It is a cluster obtained by grouping rows of a belongingness matrix of that type. It can be represented using cluster centroids.

Remark: Membership matrices T (a) are defined for objects that are connected to each other, and (b) represent objects in the same space of C dimensions.

Requirements: (a) Hidden structures across types should be consistent with each other. (b) Divergence between any two clusterings should be small

Optimization

$$\min_{W,H} \sum_{k=1}^{K} \{\|T_k - W_k H_k\|^2\} + \alpha \sum_{\substack{k=1 \\ l=1 \\ k<l}}^{K} \{\|H_k - H_l\|^2\}$$

subject to the constraints

$$W_k \geq 0 \quad \forall k = 1,2,\dots K; \qquad H_k \geq 0 \quad \forall k = 1,2,\dots K$$

Iterative Update Rules

$$W_k \leftarrow W_k \odot \frac{T_k H_k^T}{W_k H_k H_k^T} \quad \forall k = 1,2,\dots,K$$

$$H_k \leftarrow H_k \odot \frac{W_k^T T_k + \alpha \sum_{\substack{l=1 \\ k \neq l}}^{K} I^{C' \times C'} H_l}{W_k^T W_k H_k + \alpha \sum_{\substack{l=1 \\ k \neq l}}^{K} I^{C' \times C'} H_k} \quad \forall k = 1,2,\dots,K$$

$\odot$ denotes the Hadamard Product and $\frac{A}{B}$ denotes the element-wise division

Community Distribution Outlier Detection

Outlier score of an object i is the distance of the object from the nearest cluster centroid

$$OS(i) = argmin_j \; Dist(T_{k(i,)}, H_{k(j,)})$$

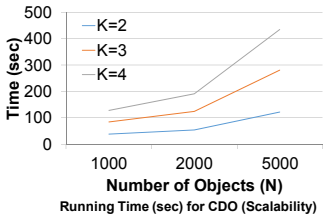
Time Complexity: Linear in number of objects.

Results on Synthetic Datasets

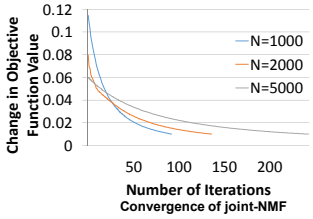
N	Outlierness (%)	|Types| = 2 CDO	SI	Homo	|Types| = 3 CDO	SI	Homo	|Types| = 4 CDO	SI	Homo
1000	1	99.5	93.0	52.0	86.0	83.3	58.0	73.0	70.8	54.5
	2	95.5	88.2	55.8	83.3	76.5	60.7	74.0	72.8	55.5
	5	75.0	72.6	58.7	66.9	67.3	57.3	67.2	67.4	56.2
2000	1	98.8	96.0	52.2	85.8	85.0	57.5	74.2	70.8	52.5
	2	97.2	85.6	57.5	81.0	75.2	55.9	73.8	67.0	57.4
	5	78.4	73.8	59.8	70.5	69.5	57.7	66.9	67.4	55.4
5000	1	99.6	99.8	56.0	84.6	83.7	54.6	76.2	76.0	53.8
	2	97.0	91.1	58.4	82.7	77.6	57.2	73.8	72.4	57.0
	5	77.2	73.3	62.0	71.9	71.2	59.5	66.2	67.6	55.4

SI (2.9%)
Homo(21%)

Synthetic Dataset Results (CDO =The Proposed Algorithm CDODA, SI = Single Iteration Baseline, Homo = Homogenous (Single NMF) Baseline) for C=6



Running Time (sec) for CDO (Scalability)



Convergence of joint-NMF

DBLP Outliers

- **Top conference outlier:** From integrated publication and information systems to virtual information and knowledge environments - Databases (0.5), Artificial Intelligence (0.09), Human Computer interaction (0.4)
- **Top terms outlier:** military - Algorithms and theory (0.02), Security and Privacy (0.37), Databases (0.22), Computer Graphics (0.37)

Summary

- Introduced outliers with respect to latent communities for heterogeneous networks.
- Proposed a joint-NMF optimization framework to learn distribution patterns across multiple object types
- Proposed an iterative two stage approach for outlier detection
- Experimented with multiple real and synthetic datasets

References

[Gao et al., 2010] Gao, J., Liang, F., Fan, W., Wang, C., Sun, Y., and Han, J. (2010). On Community Outliers and their Efficient Detection in Information Networks. KDD, page 813-822.

[Gupta et al., 2012a] Gupta, M., Gao, J., Sun, Y., and Han, J. (2012a). Community Trend Outlier Detection using Soft Temporal Pattern Mining. ECML PKDD, page 692-708.

[Gupta et al., 2012b] Gupta, M., Gao, J., Sun, Y., and Han, J. (2012b). Integrating Community Matching and Outlier Detection for Mining Evolutionary Community Outliers. KDD, page 859-867.

Microsoft

Paper ID: 60

INARC | ARL | NETWORK SCIENCE CTA