电子科技大学
University of Electronic Science and Technology of China

Paper sharing: Graph mining on 2017 kdd

Reporter: Zhongjing Yu

Data Mining Lab, Big Data Research Center, UESTC
Email：zhongjingyu@std.uestc.edu.cn
http://staff.uestc.edu.cn/yuzhongjing

Paper List

- **Clustering**
  - Local Higher-Order Graph Clustering
  - A Local Algorithm for Structure-Preserving Graph Cut
  - Graph Edge Partitioning via Neighborhood Heuristic
- **Embedding**
  - struc2vec: Learning Node Representations from Structural Identity
- **Feature selection**
  - Unsupervised Feature Selection in Signed Social Networks
- **Application**
  - TrioVecEvent: Embedding-Based Online Local Event Detection in Geo-Tagged Tweet Streams

Local Higher-Order Graph Clustering

——Jure Group

数据挖掘实验室

Data Mining Lab

REMINBER A PAPER?

《Higher-order organization of complex networks》, science 2016

Idea: a generalized framework for clustering networks on the basis **of higher-order connectivity patterns**

NETWORK SCIENCE

Higher-order organization of complex networks

Austin R. Benson,[1] David F. Gleich,[2] Jure Leskovec[3*]

Networks are a fundamental tool for understanding and modeling complex systems in physics, biology, neuroscience, engineering, and social science. Many networks are known to exhibit rich, lower-order connectivity patterns that can be captured at the level of individual nodes and edges. However, higher-order organization of complex networks—at the level of small network subgraphs—remains largely unknown. Here, we develop a generalized framework for clustering networks on the basis of higher-order connectivity patterns. This framework provides mathematical guarantees on the optimality of obtained clusters and scales to networks with billions of edges. The framework reveals higher-order organization in a number of networks, including information propagation units in neuronal networks and hub structure in transportation networks. Results show that networks exhibit rich higher-order organizational structures that are exposed by clustering based on higher-order connectivity patterns.

Networks are a standard representation of data throughout the sciences, and higher-order connectivity patterns are essential to understanding the fundamental structures that control and mediate the behavior of many complex systems (*1*–*7*). The most common higher-order structures are small network subgraphs, which we refer to as network motifs (Fig. 1A). Network motifs are considered building blocks for complex networks (*1*, *8*). For example, feed-forward loops (Fig. 1A, M_5) have proven fundamental to understanding transcriptional regulation networks (*9*); triangular motifs (Fig. 1A, M_1–M_7) are crucial for social networks (*4*); open bidirectional wedges (Fig. 1A, M_{13}) are key to structural hubs in the brain (*10*); and two-hop paths (Fig. 1A, M_8–M_{13}) are essential to understanding air traffic patterns (*5*). Although network motifs have been recognized as fundamental units of networks, the higher-order organization of networks at the level of network motifs largely remains an open question.

Here, we use higher-order network structures to gain new insights into the organization of complex systems. We develop a framework that identifies clusters of network motifs. For each network motif (Fig. 1A), a different higher-order clustering may be revealed (Fig. 1B), which means that different organizational patterns are exposed, depending on the chosen motif.

Conceptually, given a network motif M, our framework searches for a cluster of nodes S with two goals. First, the nodes in S should participate

minimizes the following ratio:

$$\phi_M(S) = \mathrm{cut}_M(S, \overline{S}) / \min[\mathrm{vol}_M(S), \mathrm{vol}_M(\overline{S})] \quad (1)$$

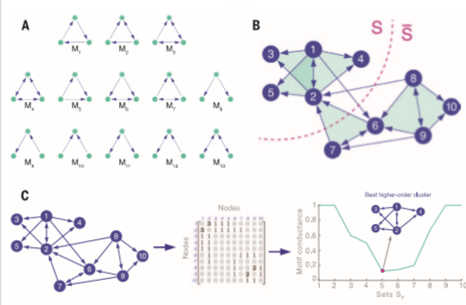
where $\overline{S}$ denotes the remainder of the nodes (the complement of S), $\mathrm{cut}_M(S, \overline{S})$ is the number of instances of motif M with at least one node in S and one in $\overline{S}$, and $\mathrm{vol}_M(S)$ is the number of nodes

in instances of M that reside in S. Equation 1 is a generalization of the conductance metric in spectral graph theory, one of the most useful graph partitioning scores (*11*). We refer to $\phi_M(S)$ as the motif conductance of S with respect to M.

Finding the exact set of nodes S that minimizes the motif conductance is computationally infeasible (*12*). To approximately minimize Eq. 1 and, hence, to identify higher-order clusters, we developed an optimization framework that provably finds near-optimal clusters [supplementary materials (*13*)]. We extend the spectral graph clustering methodology, which is based on the eigenvalues and eigenvectors of matrices associated with the graph (*11*), to account for higher-order structures in networks. The resulting method maintains the properties of traditional spectral graph clustering: computational efficiency, ease of implementation, and mathematical guarantees on the near-optimality of obtained clusters. Specifically, the clusters identified by our higher-order clustering framework satisfy the motif Cheeger inequality (*14*), which means that our optimization framework finds clusters that are at most a quadratic factor away from optimal.

The algorithm (illustrated in Fig. 1C) efficiently identifies a cluster of nodes S as follows:

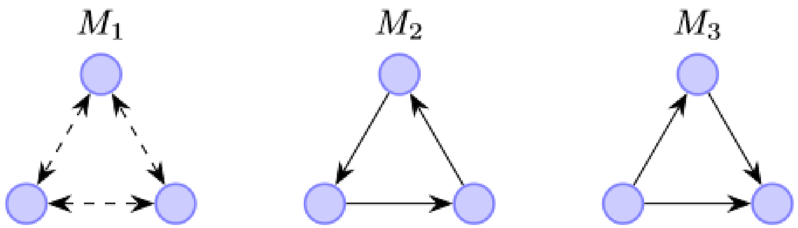
• Step 1: Given a network and a motif M of interest, form the motif adjacency matrix W_M whose entries (i, j) are the co-occurrence counts of nodes i and j in the motif M: $(W_M)_{ij}$ = number of instances of M that contain nodes i and j.

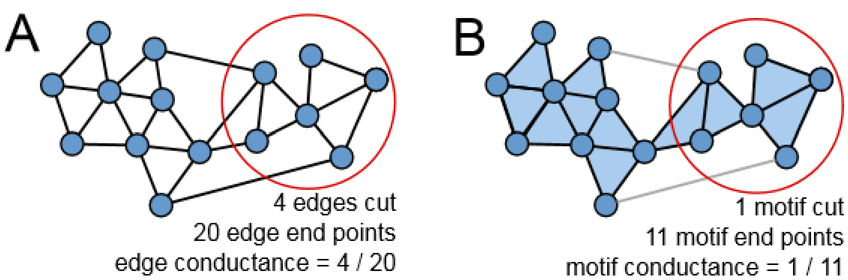


Local Higher-Order Graph Clustering
——Jure Group

数据挖掘实验室
Data Mining Lab

Idea : incorporating **higher-order network information** captured by small subgraphs, also called network motifs (**constructing new matrix**). And they develop the Motif-based **Approximate Personalized PageRank** (MAPPR) algorithm that finds clusters containing a seed node with minimal motif conductance.





Introduction of PageRank Nibble for community detection

数据挖掘实验室 Data Mining Lab

Idea: The random walker starts on an initial node and moves to a neighboring node based on the probabilities of the connecting edges. If the walker goes into a **dense region**, it would be **hard to get out of the region.**

Method:

- constructing transition matrix
- getting pagerank value(V_r) for each nodes
- if V_r /W_r > threshold for each node: collecting the node, where W_r is the weight of a node

Ref: Peng W, Wang J, Zhao B, et al. Identification of protein complexes using weighted pagerank-nibble algorithm and core-attachment structure[J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2015, 12(1): 179-192

A Local Algorithm for Structure-Preserving Graph Cut ——Jingrui He Group

Idea: traditional methods to find graph cut just consider connectivity between nodes but higher network structures. In this paper, authors focus on mining user-specified **high-order network structures** and aim to find a structure-rich subgraph which does not break many such structures by separating the subgraph from the rest.

Method: adjacent matrix ➔ **adjacent tensor**.

Pagerank to find graph cut.

$\mathbb{N}$	Example	Illustration	Markov Chain
1^{st}-order	Vertex		0^{th}-order
2^{nd}-order	Edge		1^{st}-order
3^{rd}-order	3-node Line		2^{nd}-order
	Triangle		
k^{th}-order	k-node Star		$(k-1)^{th}$-order

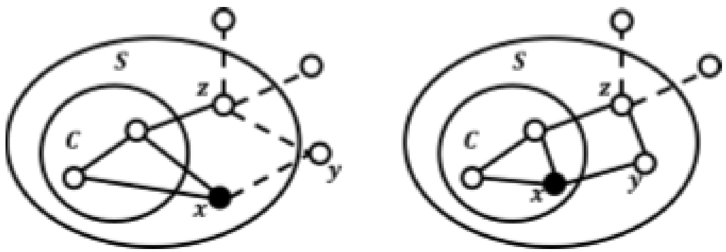
Graph Edge Partitioning via Neighborhood Heuristi

——Hong Kong University, Stanford,

Huawei Noah's Ark Lab

数据挖掘实验室
Data Mining Lab

Idea: like a process of diffusion, and **provide a worst-case upper bound** of replication factor for their heuristic on general graphs.



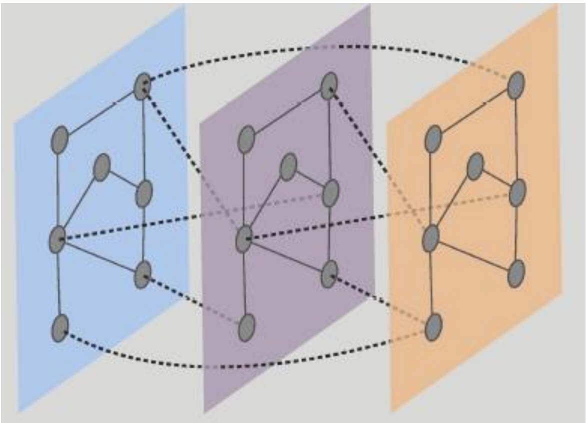
Proving : balabalabala~~~~

struc2vec: Learning Node Representations from Structural Identity

——Leonardo et. al

数据挖掘实验室
Data Mining Lab

Idea: a novel and flexible framework for learning latent representations for **the structural identity**(**structural similarity**) of nodes. *struc2vec* uses a hierarchy to measure node similarity at different scales, and constructs a **multilayer graph** to encode **structural similarities** and **generate structural context** for nodes.



A tiny example(aren't serious)

struc2vec: Learning Node Representations from Structural Identity
——Leonardo et. al

数据挖掘实验室
Data Mining Lab

Steps:

- **Measuring structural similarity**

  Two nodes that have the same degree are structurally similar, but if their neighbors also have the same degree, then they are even more structurally similar. (**DTW** $f_k(u,v)$)

- **Constructing the context graph**

  Let M denote the multilayer graph where layer k is defined using the k-hop neighborhoods of the nodes.

  same layer: $w_k(u,v) = e^{-f_k(u,v)}, \quad k = 0, \ldots, k^*$ (**undirected graph**)

  neighboring layers : $w(u_k, u_{k+1}) = \log(\Gamma_k(u) + e), \quad k = 0, \ldots, k^* - 1$ (**directed graph**)

  $w(u_k, u_{k-1}) = 1, \quad k = 1, \ldots, k^*$

  u_k, u_{k+1} is corresponding vertex in layer *k* and *k*+1

struc2vec: Learning Node Representations from Structural Identity

——Leonardo et. al

数据挖掘实验室

Data Mining Lab

➢ Constructing the context graph

$$\Gamma_k(u) = \sum_{v \in V} \mathbb{1}(w_k(u,v) > \overline{w_k})$$

Where $\Gamma_k(u)$ is number of edges incident to u that have weight larger than the average edge weight of the complete graph in layer k. **Note that** if u **has many similar nodes in the current layer**, then it should **change layers** to obtain a **more refined context**.

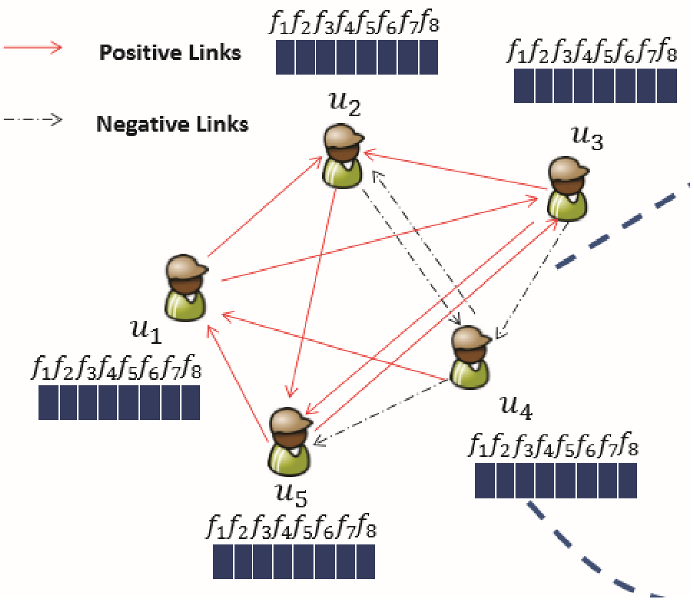
➢ Generating context for nodes

Random walk: start a node in layer 0. Random walks **have a fixed** and relatively **short length** (number of steps), and the process is repeated a certain number of times, giving rise to multiple independent walks. Finally, the context is generated by the process.

Unsupervised Feature Selection in Signed Social Networks ——Kewei Cheng et. al

Scenario： nodes with features(attribution) are connected by positive link and negative link. How to select ?

Idea: these latent representations encode the signed network structure which selected feature should preserve. **In my word, the relationship between features and topology are consist. Then, the node latent representations can guide feature selection**



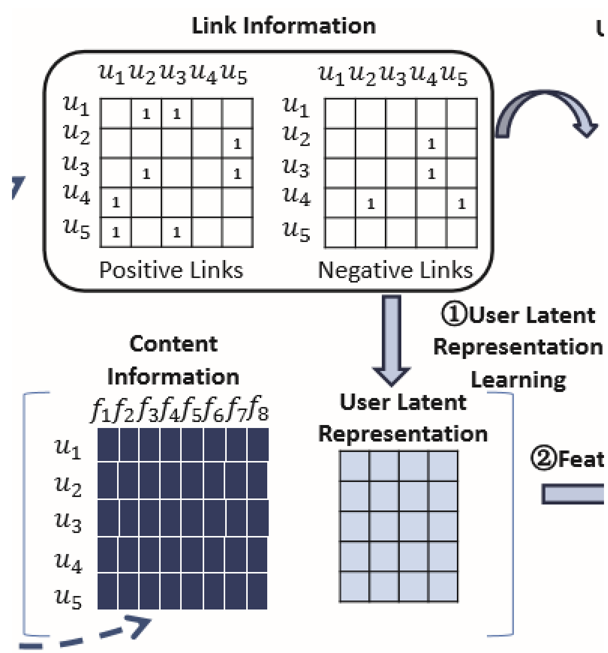
Unsupervised Feature Selection in Signed Social Networks ——Kewei Cheng et. al

Step 1: collectively factorizing A^p and A^n into a unified **low-rank** representation U

$$\min_{\mathbf{U},\mathbf{V}^p,\mathbf{V}^n} \beta^{+}\|\mathbf{O}^p \odot (\mathbf{A}^p - \mathbf{U}\mathbf{V}^p\mathbf{U}')\|_F^2 + \beta^{-}\|\mathbf{O}^n \odot (\mathbf{A}^n - \mathbf{U}\mathbf{V}^n\mathbf{U}')\|_F^2,$$

$$\mathbf{O}^p_{ij} = \begin{cases} 1, & \text{if } \mathbf{A}^p_{ij} = 1 \\ 0, & \text{otherwise} \end{cases}$$

$$\mathbf{O}^n_{ij} = \begin{cases} 1, & \text{if } \mathbf{A}^n_{ij} = 1 \\ 0, & \text{otherwise} \end{cases}$$

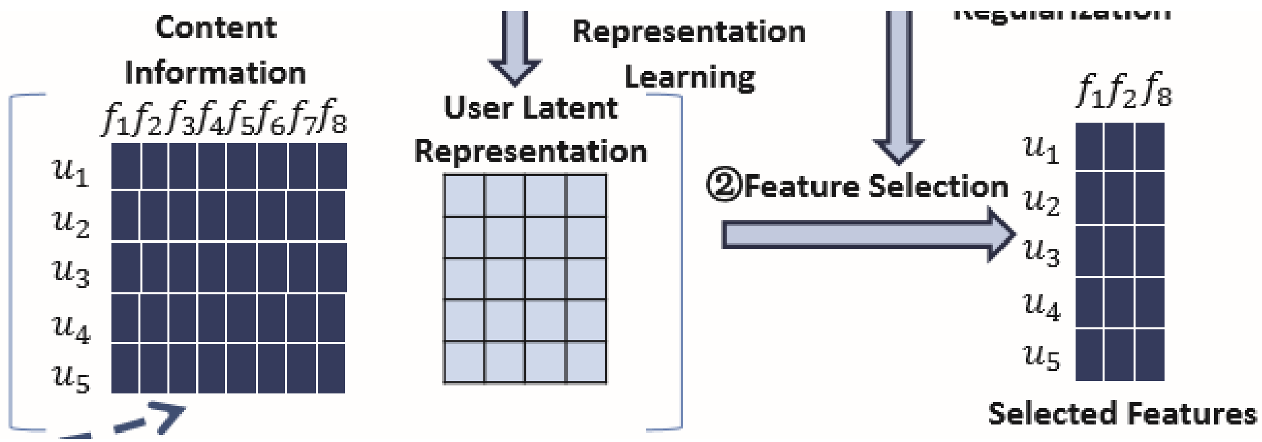


Unsupervised Feature Selection in Signed Social Networks ——Kewei Cheng et. al

Step 2: leveraging the user latent representations U to guide feature selection via a multivariate linear regression model.

$$\min_{\mathbf{W}} \|\mathbf{XW} - \mathbf{U}\|_F^2 + \alpha\|\mathbf{W}\|_{2,1}$$

Where X is attributions(features).



Unsupervised Feature Selection in Signed Social Networks ——Kewei Cheng et. al

数据挖掘实验室
Data Mining Lab

Step 3: modeling user proximity(***omitting***)

Finally, object function:

$$\min_{\mathbf{W},\mathbf{U},\mathbf{V}^p,\mathbf{V}^n} \|\mathbf{X}\mathbf{W}-\mathbf{U}\|_F^2 + \alpha\|\mathbf{W}\|_{2,1} + \frac{\gamma}{2}tr(\mathbf{U}'\mathbf{L}\mathbf{U})$$
$$+\frac{\beta^+}{2}\|\mathbf{O}^p \odot (\mathbf{A}^p - \mathbf{U}\mathbf{V}^p\mathbf{U}')\|_F^2$$
$$+\frac{\beta^-}{2}\|\mathbf{O}^n \odot (\mathbf{A}^n - \mathbf{U}\mathbf{V}^n\mathbf{U}')\|_F^2,$$

TrioVecEvent: Embedding-Based Online Local Event Detection in Geo-Tagged Tweet Streams

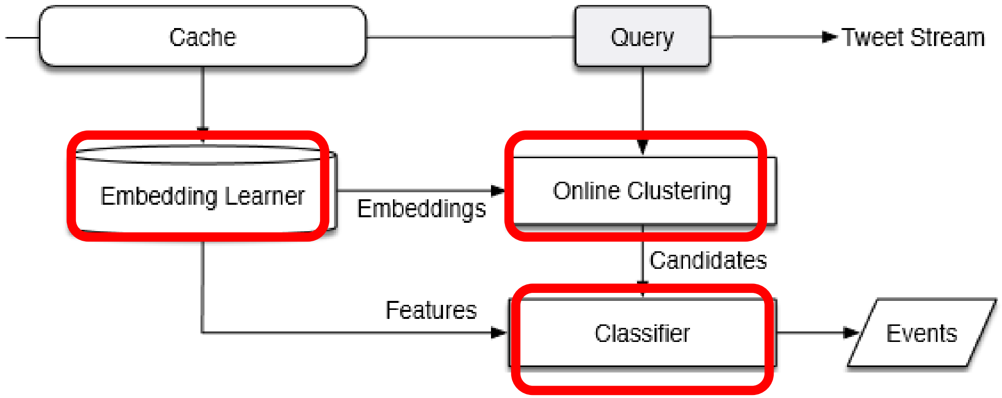
——Jiawei Han Group

数据挖掘实验室

Data Mining Lab

Question: online local events detection by *geo-tagged* tweet stream.

Method:



Embedding learner: map all the regions, hours, and keywords into sample space.

Online clustering: Bayesian mixture model

Classifier: detection events

TrioVecEvent: Embedding-Based Online Local Event Detection in Geo-Tagged Tweet Streams

——Jiawei Han Group

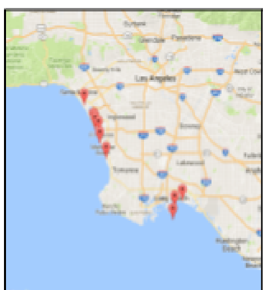
数据挖掘实验室
Data Mining Lab

Part 1: embedding learner :

- Capturing the semantic similarities between tweets and further group tweets.
- Revealing keywords appearing in different regions and hours (background knowledge)

Idea:

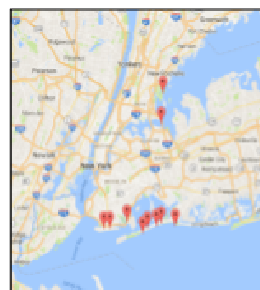
1. Discretization(regions, hours, and keywords).



"beach"

lax	lakers	dodgers	beachlife
international	kobe	ladders	sand
losangeles	bryant	dogerstadium	boardwalk
united	bulls	itfdb	ocean
people	cavs	letsgododgers	wave
tsa	kevin	game	beachday
sfo	knicks	dodgergame	pacificocean
food	clipper	play	santamonica
flight	lebron	losdoyers	pier
travel	cp3	win	wave
"33.942, -118.409"	"nba"	"baseball"	"beach"

(a) Examples on LA (the second query is the location of the LAX Airport).



"beach"

jfk	knicks	mlb	rockaway
airport	melo	yankees	beachday
international	lebron	mets	howard_beach
johnfkennedy	durant	yanks	brighton
burger	basketball	inning	longbeach
terminal	kobe	yankee	coney
john	cavs	ballpark	atlantic
kennedy	theknicks	pitch	island
sfo	game	jeter	boardwalk
flight	lakers	game	long
"40.641, -73.778"	"nba"	"baseball"	"beach"

(b) Examples on NY (the second query is the location of the JFK Airport).

TrioVecEvent: Embedding-Based Online Local Event Detection in Geo-Tagged Tweet Streams

——Jiawei Han Group

数据挖掘实验室
Data Mining Lab

Idea:

2. learning embedding by **Continuous Bag of Words Model（CBOW）[**predicting one unit given its context**]**.

Method: given a tweet d, for any unit i, let v_i be the embedding of unit i, then we model the likelihood J_C

$$p(i|d_{-i}) = \exp\big(s(i, d_{-i})\big) / \Sigma_{j \in X} \exp(s(j, d_{-i}))$$

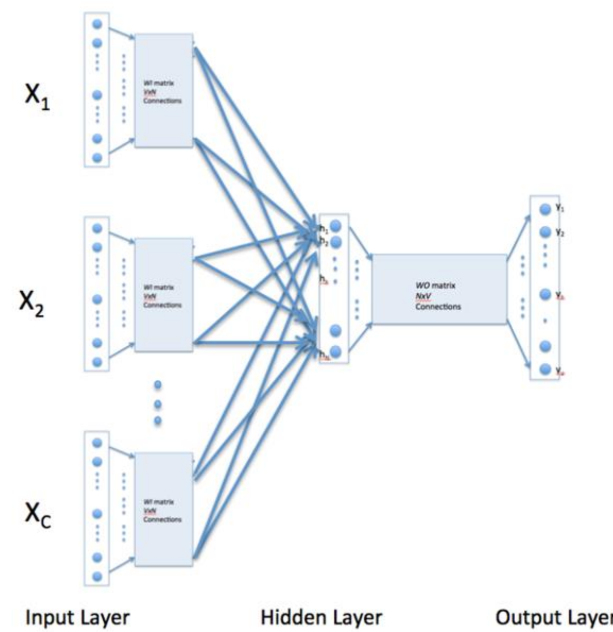
——computing the probability of words.

$$s(i, d_{-i}) = \mathbf{v}_i^T \Sigma_{d \in d_{-i}} \mathrm{v}_j / |d_{-i}|$$ —— similarity

$$J_C = -\Sigma_{d \in C} \Sigma_{i \in d} \log p(i\,|d_{-i})$$——likelihood

$$J_d = -log\sigma\big(s(i, d_{-i})\big) - \Sigma_{k=1}^{K} log\sigma\big(-s(k, d_{-i})\big)$$

——cross entropy



Training the neural network by min cross entropy

TrioVecEvent: Embedding-Based Online Local Event Detection in Geo-Tagged Tweet Streams

——Jiawei Han Group

数据挖掘实验室
Data Mining Lab

Part 2: online clustering: Bayesian mixture model

Basic idea: every **geo-topic cluster** implies a **coherent activity** (e.g., protest) around a certain geo-location (e.g., the JFK Airport). location acts as a geographical center that triggers geo-location observations around it in the Euclidean space; while the **activity serves** as a **semantic focus** that triggers **semantic embedding observations around it** in the spherical space.

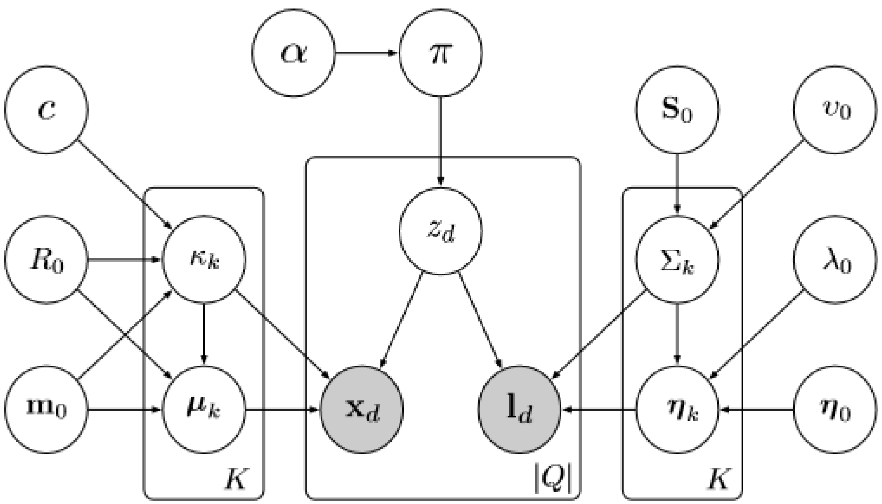
TrioVecEvent: Embedding-Based Online Local Event Detection in Geo-Tagged Tweet Streams

——Jiawei Han Group

数据挖掘实验室
Data Mining Lab

Station of variable and data formation

Data formation : each tweet d as a tuple (I_d, x_d), where I_d is location, x_d is the D-dimensional semantic embedding of d

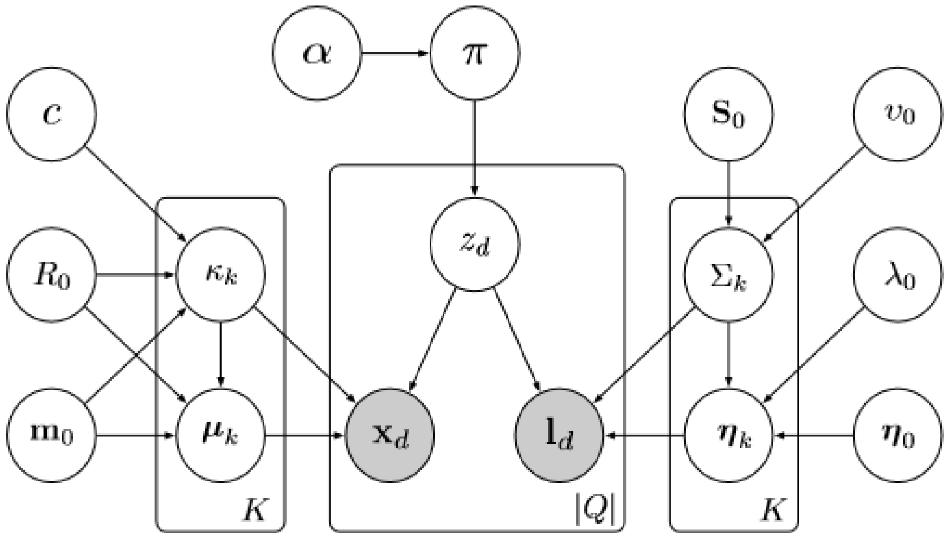


$\mathcal{X}$	the set of semantic embeddings for the tweets in Q
$\mathcal{Z}$	the set of cluster memberships for the tweets in Q
$\mathcal{L}$	the set of geo-location vectors for the tweets in Q
$\boldsymbol{\kappa}$	the set of κ for all the clusters
$\boldsymbol{\kappa}^{\neg k}$	the subset of $\boldsymbol{\kappa}$ excluding the one for cluster k
$\mathrm{A}^{\neg d}$	the subset of any set A excluding element d
A^{k}	the subset of elements that are assigned to cluster k in set A
$\mathbf{x}^{k}$	the sum of the semantic embeddings in cluster k
$\mathbf{x}^{k,\neg d}$	the sum of the semantic embeddings in cluster k excluding d
n^{k}	the number of tweets in cluster k
$n^{k,\neg d}$	the number of tweets in cluster k excluding d

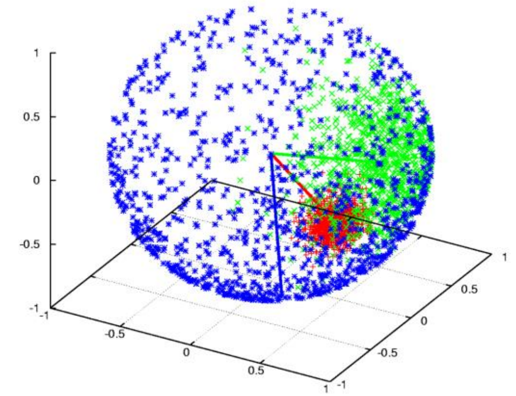
TrioVecEvent: Embedding-Based Online Local Event Detection in Geo-Tagged Tweet Streams

——Jiawei Han Group

数据挖掘实验室
Data Mining Lab



$$\pi \sim \text{Dirichlet}(.|\alpha)$$
$$\{\eta_k, \Sigma_k\} \sim \text{NIW}(.|\eta_0, \lambda_0, S_0, \upsilon_0) \quad k = 1, 2, \ldots, K$$
$$\{\mu_k, \kappa_k\} \sim \Phi(.|m_0, R_0, c) \quad k = 1, 2, \ldots, K$$
$$z_d \sim \text{Categorical}(.|\pi) \quad d \in Q$$
$$l_d \sim \mathcal{N}(.|\eta_{z_d}, \Sigma_{z_d}) \quad d \in Q$$
$$x_d \sim \text{vMF}(.|\mu_{z_d}, \kappa_{z_d}) \quad d \in Q$$



vMF from [Wiki]()

TrioVecEvent: Embedding-Based Online Local Event Detection in Geo-Tagged Tweet Streams

——Jiawei Han Group

数据挖掘实验室
Data Mining Lab

Part 3 :Classifier: detection events

- **Spatial unusualness** quantifies how unusual a candidate is in its geographical region.
- **Temporal unusualness** quantifies how temporally unusual a candidate is
- **Spatiotemporal unusualness** jointly considers the space and time to quantify how unusual a candidate is.
- **Semantic concentration** computes how semantically coherent is.
- **Spatial and temporal concentrations** quantify how concentrated a candidate C is over the space and time.
- **Burstiness** quantifies how bursty a candidate C is.

Finally, they train a binary classifier and judge whether each candidate is indeed a local event

Thanks

Zhongjing Yu
yuzhongjing@std.uestc.cn