

Summary: Chapter 8-9: Classification

- Concepts, supervised learning, and applications
- Classification models
 - Decision Tree (ID₃, C_{4.5}, CART)
 - Information gain, gain ratio, gini index
 - Naïve Bayes
 - Support Vector Machines
 - Artificial Neural Networks, Deep Neural Networks
- Classification evaluation
 - Accuracy, precision, recall, F₁, ROC curve, P@K...

Course Project

- **Data objects:** papers, authors ...
- **Labels:** conferences
- **Features:** paper's terms, author's paper's terms ...
- **Models:** decision trees (ID3, C4.5, CART), naïve bayes, support vector machines (LinearSVM, KernelSVM), neural networks

NIPS Challenges 2017

- <https://www.kaggle.com/c/msk-redefining-cancer-treatment>
- Data objects?
- Labels?
- Features?
- Models?

Schedule

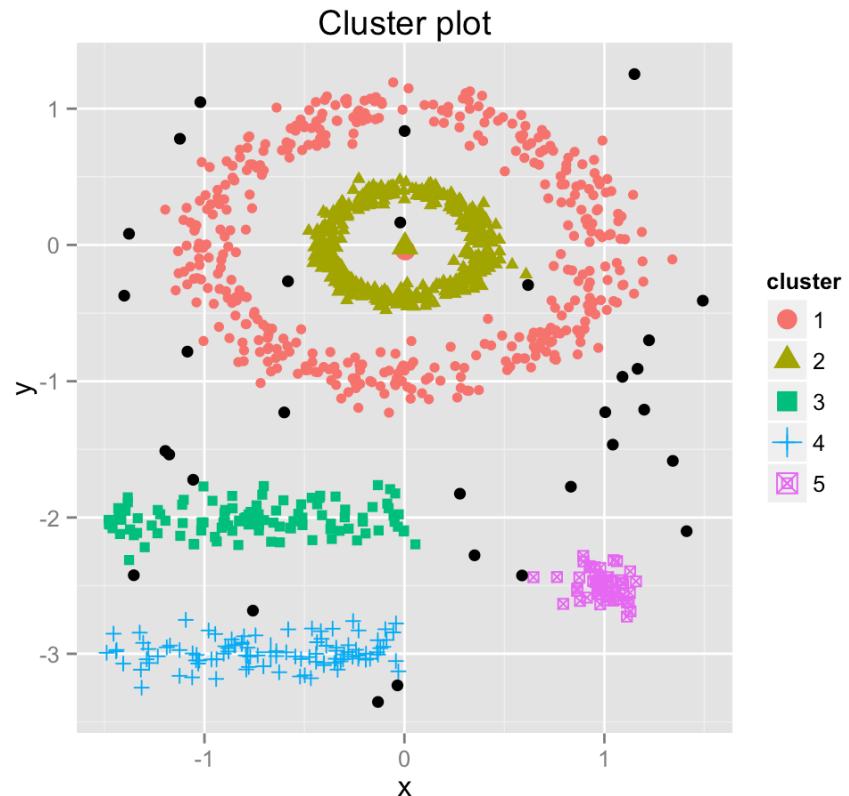
- Nov. 9: **HW4 due**, HW5 out
- Nov. 21: Last lecture of new content
- Thanksgiving break
- Nov. 28: Course review, **HW5 due**
- Nov. 30: Course review, **Project due**
 - Early project presentations (+ $10 \times 25\%$ points)
 - Let me know before/at Nov. 28's class (email).
 - Submission! You can update till Final exam (Nov. 12)
- Dec. 5 and Dec. 7: Project presentations

Chapter 10. Cluster Analysis: Concepts

Meng Jiang

CSE 40647/60647 Data Science Fall 2017

Introduction to Data Mining

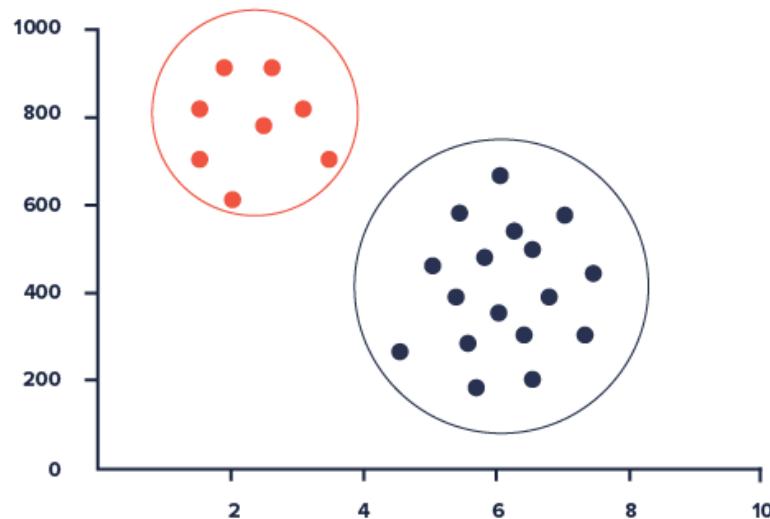


Outline

- **Cluster, Cluster analysis, Applications**
- “Good” cluster, Quality function, Similarity measure
- Considerations for Cluster Analysis, Properties of Clustering
- Requirements, Challenges
- Typical methodologies

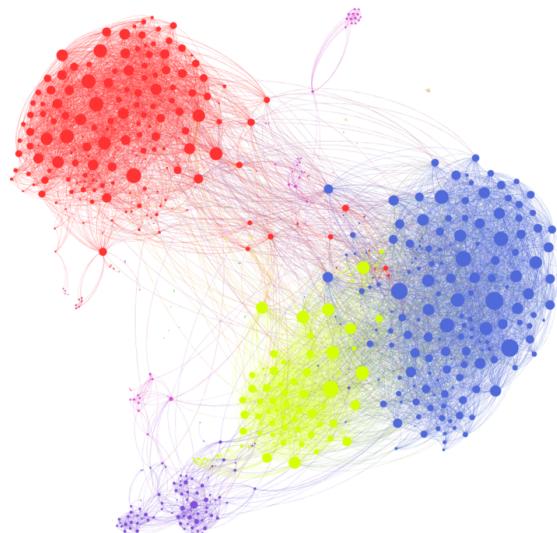
What is Cluster?

- A **cluster** is a collection of data objects which are
 - **Similar** (or related) to one another **within** the same group (i.e., cluster)
 - **Dissimilar** (or unrelated) to the objects **in other** groups (i.e., clusters)



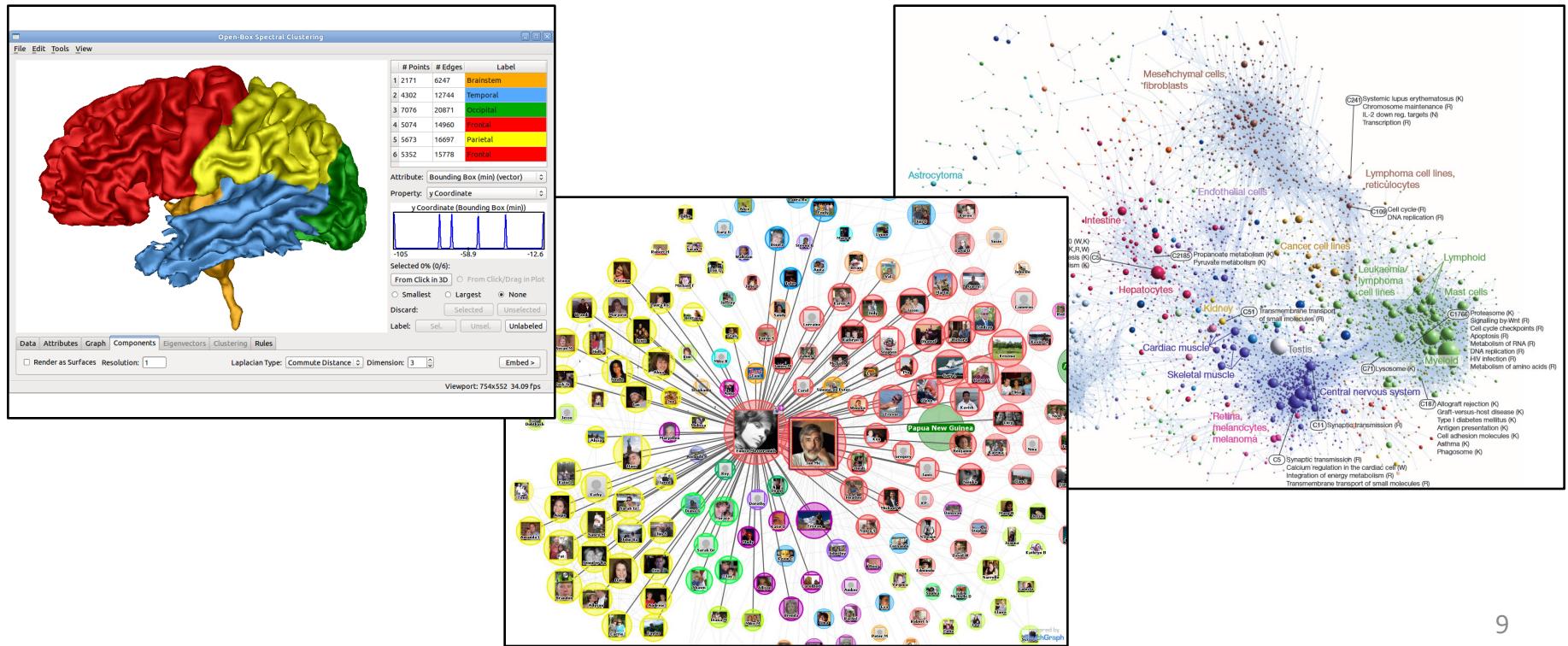
What is Cluster Analysis?

- Cluster Analysis (Clustering)
 - Def. 1. Given a set of data points, partition them into a set of clusters which are as **similar** as possible.
 - Def. 2. Grouping data objects based **only** on information found in the data describing these objects and their relationships (i.e., object/features or node/links).



Applications

- Typical ways to use/apply cluster analysis
 - As a stand-alone tool to get insight into data distribution
 - As a preprocessing/intermediate step for other algorithms



Applications (cont.)

- A key intermediate step for other data mining tasks
 - Generating a **compact summary** of data for classification, pattern discovery, hypothesis generation and testing, etc.
 - **Outlier detection:** Outliers—those “far away” from any cluster
- Data **summarization, compression, and reduction**
- Dynamic trend detection
 - Clustering stream data and detecting **trends and patterns**
- **Collaborative filtering**, recommendation systems, or customer segmentation
 - Find like-minded users or similar products
- **Multimedia data analysis**, biological data analysis and social network analysis
 - Ex. Clustering images or video/audio clips, gene/protein sequences

Formalizing Clustering

Let D denote a dataset containing N observations

$$D = \{\mathbf{x}_i \mid i = 1, 2, \dots, N\}$$

where each \mathbf{x}_i corresponds to the set of features or explanatory variables – **discrete** or **continuous** (**categorical** or **numerical**) – of the i -th observation.

Then **clustering** is the task of learning a mapping of each feature set \mathbf{x} into a previously undefined grouping.

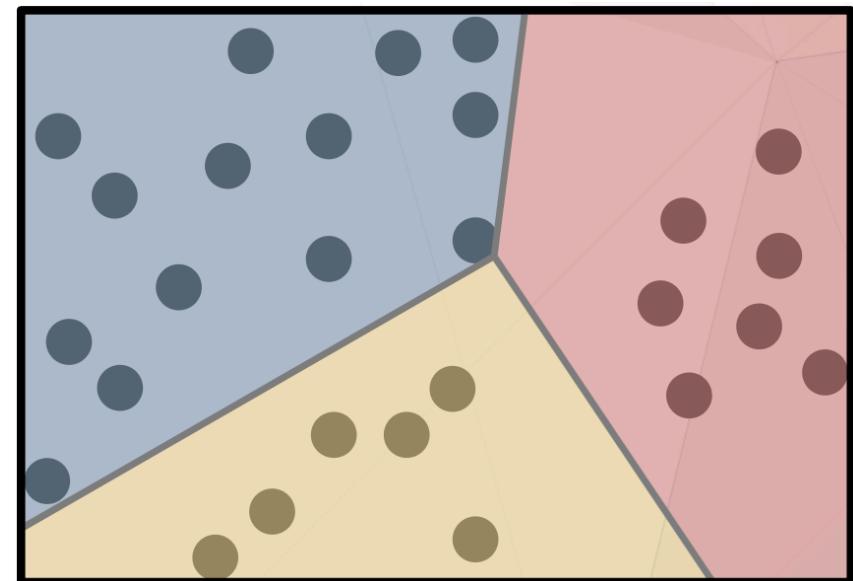
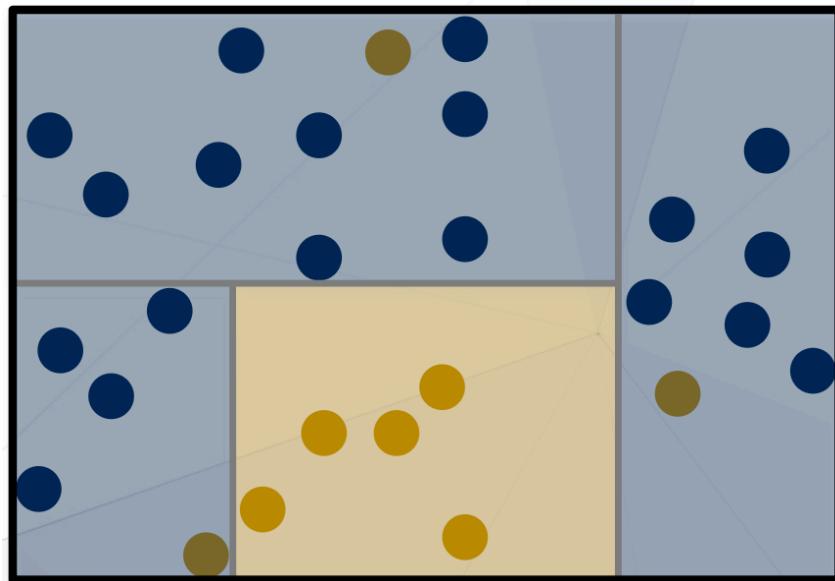
Classification vs. Clustering

Classification:

Predicts instances classes from pre-defined set of classes.

Clustering:

Finds “natural” grouping of instances with out class data.



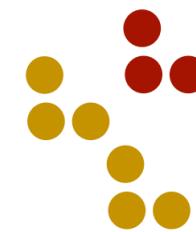
Unsupervised Learning

- Clustering
 - Input **set of features X** that can take any set of values (categorical or numerical). The goal is to identify a useful organization or grouping of the instances.
- Frequent pattern/association rule mining
 - Input **set of items X** . The goal is to discover a set of regularities, frequent itemsets, or rules between occurrences of items in the database or dataset.

Outline

- Cluster, Cluster analysis, Applications
- **“Good” cluster, Quality function, Similarity measure**
- Considerations for Cluster Analysis, Properties of Clustering
- Requirements, Challenges
- Typical methodologies

What Defines a Good Cluster?



1-cluster

2-cluster

3-cluster

4-cluster

Good Clusters

- A good clustering method will produce high quality clusters which should have
 - **High intra-class similarity:** Cohesive within clusters
 - **Low inter-class similarity:** Distinctive between clusters
- Quality function
 - There is usually a quality function that measures the “**goodness**” of a cluster
 - Depends on **similarity measure**
 - It is hard to define “similar enough” or “good enough”
 - The answer is typically highly **subjective**
 - Measured by ability to discover **hidden patterns**

Similarity and Dissimilarity

- There exist many similarity measures and/or functions for different applications
- Similarity measure is critical for cluster analysis
- Similarity
 - Measure of “alikeness” of instances
 - Greater if more alike
- Dissimilarity
 - Measure of “unalikeness” of instances
 - Lower if more alike
- Can be expressed as distance function

Outline

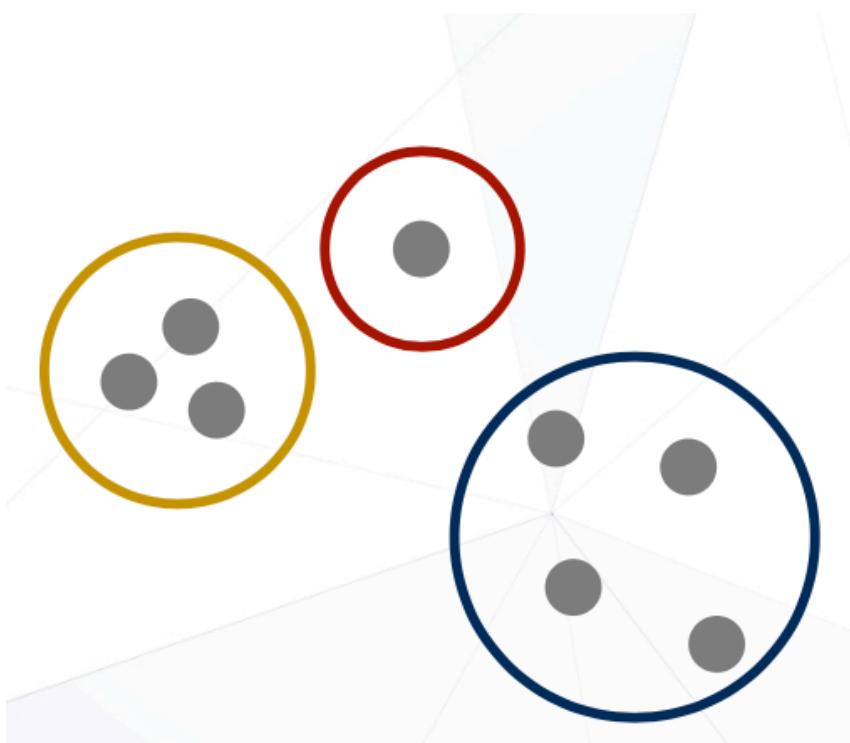
- Cluster, Cluster analysis, Applications
- “Good” cluster, Quality function, Similarity measure
- **Considerations for Cluster Analysis, Properties of Clustering**
- Requirements, Challenges
- Typical methodologies

Considerations for Cluster Analysis

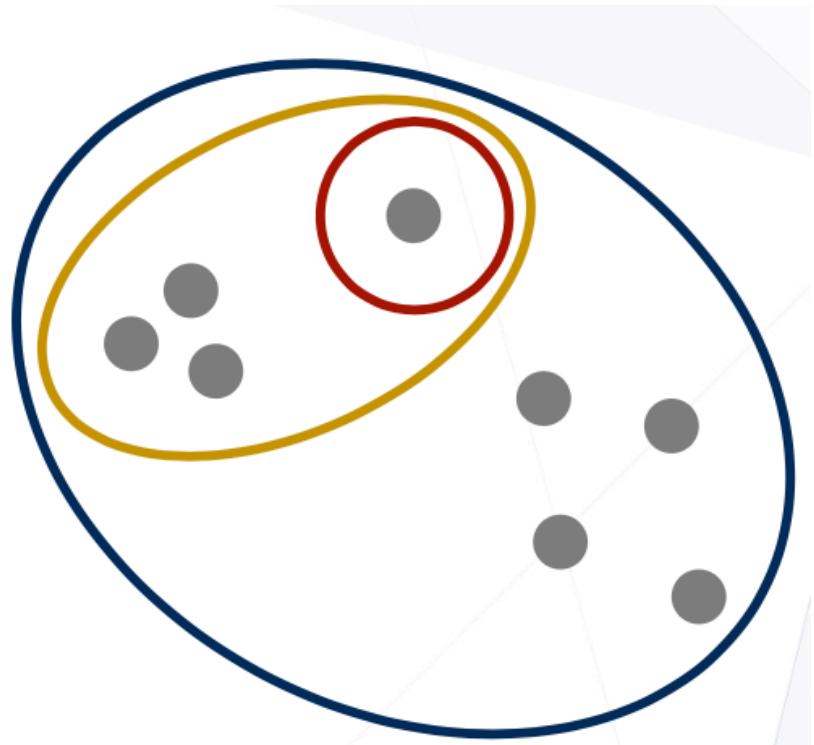
- **Partitioning criteria**
 - Single level vs. hierarchical partitioning (often, multi-level hierarchical partitioning is desirable, e.g., grouping topical terms)

Clustering Property 1

Partitional clustering



Hierarchical clustering

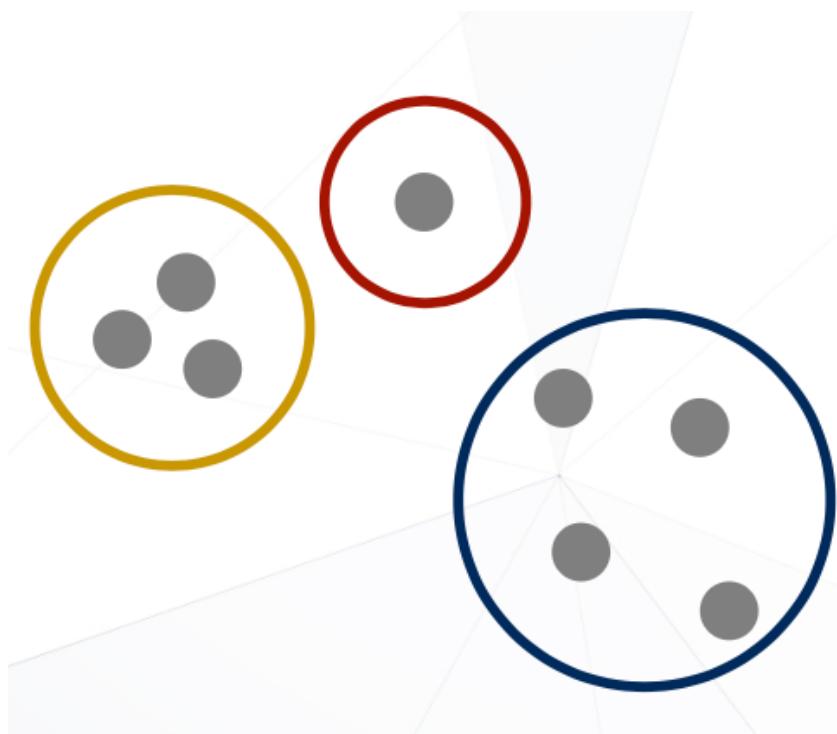


Considerations for Cluster Analysis

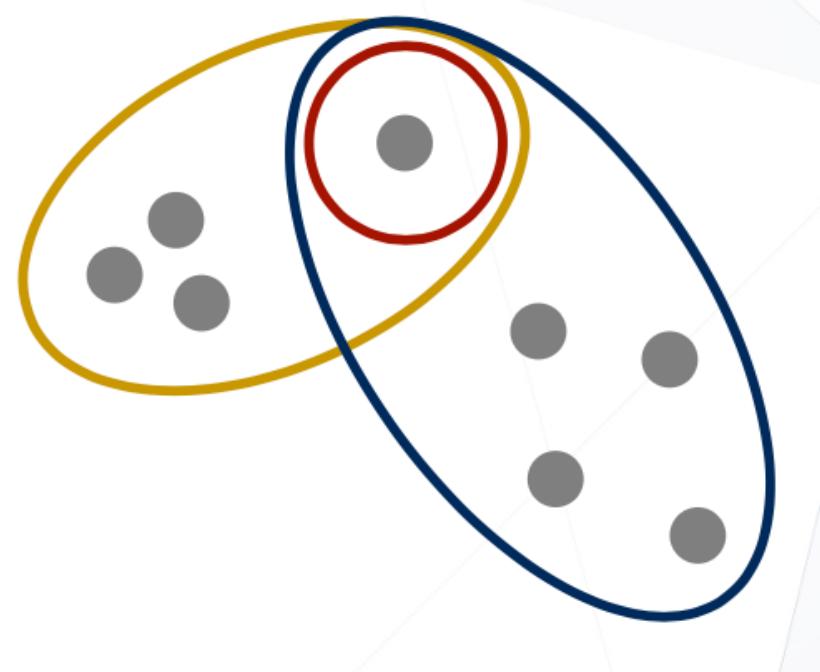
- **Partitioning criteria**
 - Single level vs. hierarchical partitioning (often, multi-level hierarchical partitioning is desirable, e.g., grouping topical terms)
- **Separation of clusters**
 - Exclusive (e.g., one customer belongs to only one region) vs. non-exclusive (e.g., one document may belong to more than one class)

Clustering Property 2

Exclusive clustering



Overlapping clustering

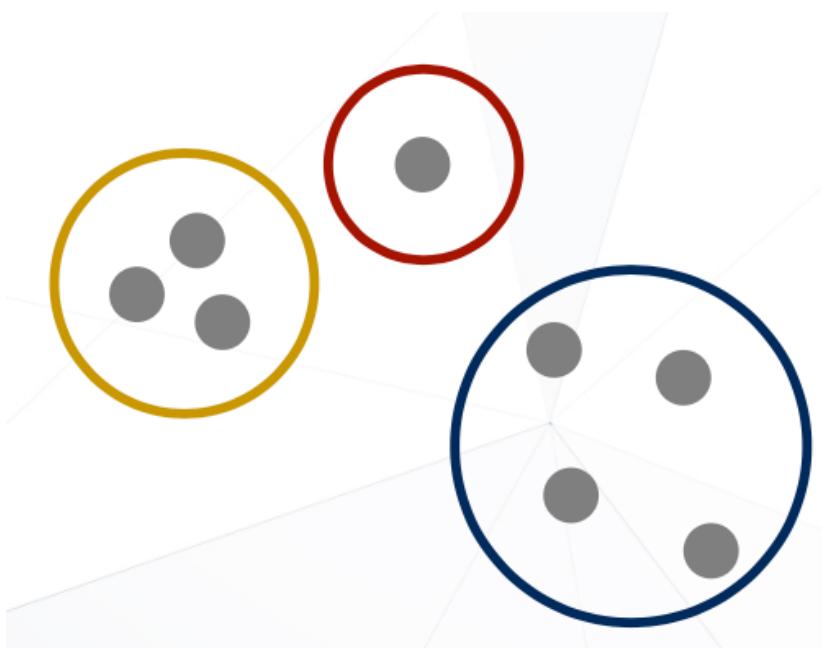


Considerations for Cluster Analysis

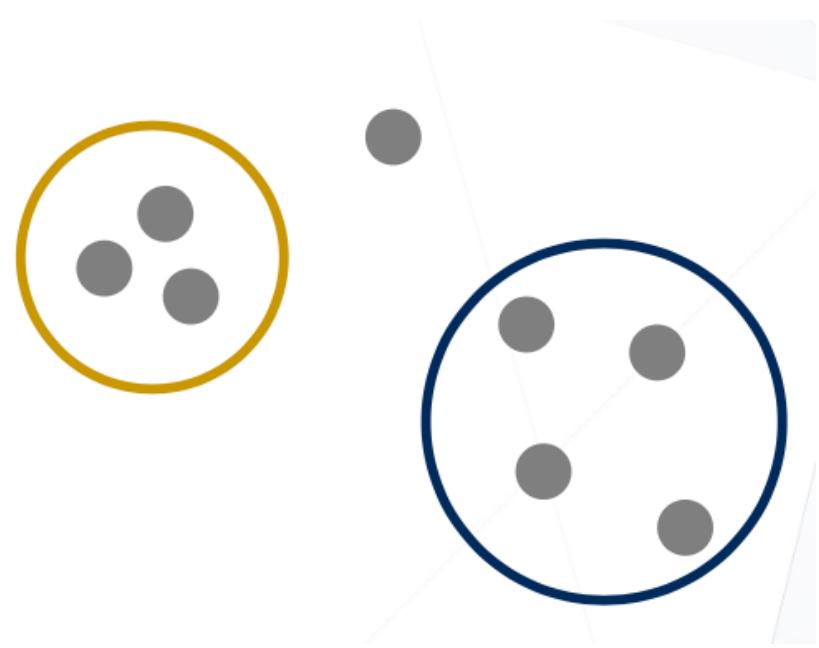
- **Partitioning criteria**
 - Single level vs. hierarchical partitioning (often, multi-level hierarchical partitioning is desirable, e.g., grouping topical terms)
- **Separation of clusters**
 - Exclusive (e.g., one customer belongs to only one region) vs. non-exclusive (e.g., one document may belong to more than one class)
- **Outlier vs. Cluster**
 - Complete (e.g., every item has one cluster) vs. partial (e.g., one item may not belong to any cluster)

Clustering Property 3

Complete clustering

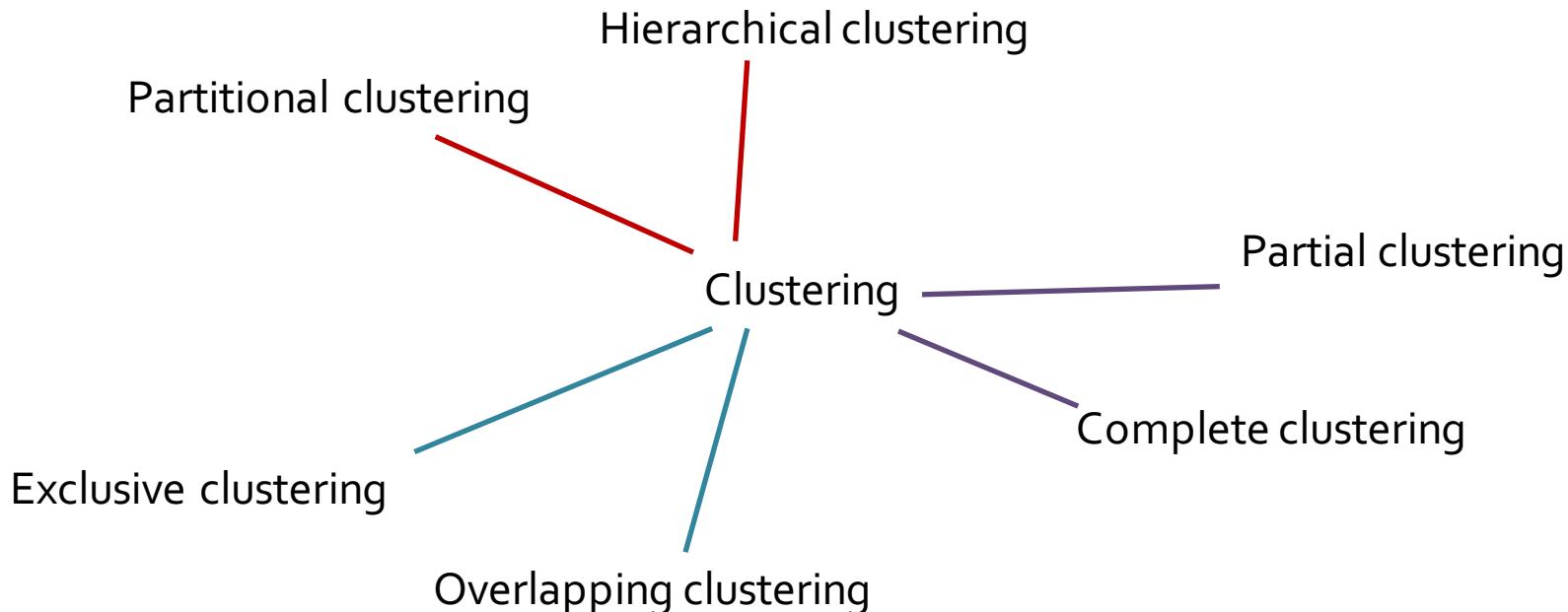


Partial clustering



Considerations → Cluster Properties

- Partitional Clustering vs. Hierarchical Clustering
- Exclusive Clustering vs. Overlapping Clustering
- Complete Clustering vs. Partial Clustering



Considerations for Cluster Analysis

- **Partitioning criteria**
 - Single level vs. hierarchical partitioning (often, multi-level hierarchical partitioning is desirable, e.g., grouping topical terms)
- **Separation of clusters**
 - Exclusive (e.g., one customer belongs to only one region) vs. non-exclusive (e.g., one document may belong to more than one class)
- **Outlier vs. Cluster**
- **Similarity measure**
 - Distance-based (e.g., Euclidean, road network, vector) vs. connectivity-based (e.g., density or contiguity)
- **Clustering space**
 - Full space (often when low dimensional) vs. subspaces (often in high-dimensional clustering)

Outline

- Cluster, Cluster analysis, Applications
- “Good” cluster, Quality function, Similarity measure
- Considerations for Cluster Analysis, Properties of Clustering
- **Requirements, Challenges**
- Typical methodologies

Requirements and Challenges

- **Quality**
 - Ability to deal with **different types of attributes**: Numerical, categorical, text, multimedia, networks, and mixture of multiple types
 - Discovery of clusters with arbitrary shape
 - Ability to deal with noisy data

Q: List some different types of data for cluster analysis?

Clustering Different Types of Data (I)

- **Numerical data**
 - Most earliest clustering algorithms were designed for numerical data
- **Categorical data** (including binary data)
 - Discrete data, no natural order (e.g., sex, race, zip-code, and market-basket)
- **Text data:** Popular in social media, Web, and social networks
 - Features: High-dimensional, sparse, value corresponding to word frequencies
 - Methods: Topic modeling
- **Multimedia data:** Image, audio, video (e.g., on Flickr, YouTube)
 - Multi-modal (often combined with text data)
 - Contextual: Containing both behavioral and contextual attributes
 - Images: Position of a pixel represents its context, value represents its behavior
 - Video and music data: Temporal ordering of records represents its meaning

Clustering Different Types of Data (II)

- **Time-series data:** Sensor data, stock markets, temporal tracking, forecasting, etc.
 - Data are temporally dependent
 - Time: contextual attribute; data value: behavioral attribute
 - Correlation-based online analysis (e.g., online clustering of stock to find stock tickers)
 - Shape-based offline analysis (e.g., cluster ECG based on overall shapes)
- **Sequence data:** Weblogs, biological sequences, system command sequences
 - Contextual attribute: Placement (rather than time)
 - Similarity functions: Hamming distance, edit distance, longest common subsequence
 - Sequence clustering: Suffix tree; generative model (e.g., Hidden Markov Model)
- **Stream data:**
 - Real-time, evolution and concept drift, single pass algorithm
 - Create efficient intermediate representation, e.g., micro-clustering

Clustering Different Types of Data (III)

- **Graphs and homogeneous networks**
 - Every kind of data can be represented as a graph with similarity values as edges
 - Methods: Generative models; combinatorial algorithms (graph cuts); spectral methods; non-negative matrix factorization methods
- **Heterogeneous networks**
 - A network consists of multiple typed nodes and edges (e.g., bibliographical data)
 - Clustering different typed nodes/links together (e.g., NetClus)
- **Uncertain data:** Noise, approximate values, multiple possible values
 - Incorporation of probabilistic information will improve the quality of clustering
- **Big data:** Model systems may store and process very big data (e.g., weblogs)
 - Ex. Google's MapReduce framework
 - Use *Map* function to distribute the computation across different machines
 - Use *Reduce* function to aggregate results obtained from the *Map* step

Requirements and Challenges

- **Quality**
 - Ability to deal with different types of attributes: Numerical, categorical, text, multimedia, networks, and mixture of multiple types
 - Discovery of clusters with arbitrary shape
 - Ability to deal with noisy data
- **Scalability**
 - Clustering all the data instead of only on samples
 - High dimensionality
 - Incremental or stream clustering and insensitivity to input order

Requirements and Challenges

- **Quality**
 - Ability to deal with different types of attributes: Numerical, categorical, text, multimedia, networks, and mixture of multiple types
 - Discovery of clusters with arbitrary shape
 - Ability to deal with noisy data
- **Scalability**
 - Clustering all the data instead of only on samples
 - High dimensionality
 - Incremental or stream clustering and insensitivity to input order
- **Constraint-based clustering**
 - User-given preferences or constraints; domain knowledge; user queries

Requirements and Challenges

- **Quality**
 - Ability to deal with different types of attributes: Numerical, categorical, text, multimedia, networks, and mixture of multiple types
 - Discovery of clusters with arbitrary shape
 - Ability to deal with noisy data
- **Scalability**
 - Clustering all the data instead of only on samples
 - High dimensionality
 - Incremental or stream clustering and insensitivity to input order
- **Constraint-based clustering**
 - User-given preferences or constraints; domain knowledge; user queries
- **Interpretability and usability**

Outline

- Cluster, Cluster analysis, Applications
- “Good” cluster, Quality function, Similarity measure
- Considerations for Cluster Analysis, Properties of Clustering
- Requirements, Challenges
- **Typical methodologies**

Typical Clustering Methodologies

- Distance-based methods
 - **Partitioning algorithms: K-Means, K-Medians, K-Medoids**
 - Hierarchical algorithms: Agglomerative vs. divisive methods
- Density-based and grid-based methods
 - **Density-based: Data space is explored at a high-level of granularity and then post-processing to put together dense regions into an arbitrary shape (DBSCAN)**
 - Grid-based: Individual regions of the data space are formed into a grid-like structure

Typical Clustering Methodologies

- Probabilistic generative models: Modeling data from a generative process
 - Assume a specific form of the generative model (e.g., mixture of Gaussians)
 - Model parameters are estimated with the Expectation-Maximization (EM) algorithm (using the available dataset, for a maximum likelihood fit)
 - Then estimate the generative probability of the underlying data points

The Ground Truth About Metadata and Community Detection in Networks

Thursday, October 26, 2017
3:30-4:45 p.m.
140 DeBartolo

Community detection is one of the most common tasks in network analysis, in which we seek to infer the underlying structural modules or groups of a network from the pattern of which nodes are connected. The standard evaluation metric of these algorithms is based on how closely these inferred communities correlate with node "metadata", i.e., node labels like a person's ethnicity in a social network or the brain region in a connectome.

In this talk, I will present two strong results on community detection and node metadata.

First, I'll introduce the No Free Lunch theorem for community detection, which proves that every community detection algorithm has the same average performance, across all network inputs, and the no-bijection theorem, which proves that no algorithm can always recover "ground truth" communities. However, by using node metadata to guide the community detection process, rather than as an evaluation target, better inferences about the network's organization can be obtained. To substantiate this claim, I'll introduce a Bayesian stochastic block model that automatically learns the correlation between node metadata and network communities, if any exists. The learned correlations are interesting in their own right, and allow us to make predictions about the community membership of nodes whose network connections are unknown. This method performs better than any algorithm that uses only structure or only metadata alone, and I will demonstrate its application to several real-world networks drawn from social, biological, and technological domains.

This is joint work with Leto Peel, Daniel B. Larremore, and Mark Newman.

Aaron Clauset is an Assistant Professor in the Department of Computer Science and the BioFrontiers Institute at the University of Colorado Boulder, and is External Faculty at the Santa Fe Institute. He received a PhD in Computer Science, with distinction, from the University of New Mexico, a BS in Physics, with honors, from Haverford College, and was an Omidyar Fellow at the prestigious Santa Fe Institute.

Clauset is an internationally recognized expert on network science, data science, and machine learning for complex systems. His work has appeared in many prestigious scientific venues, including Nature, Science, PNAS, JACM, WWW, ICWSM, STOC, SIAM Review, and Physical Review Letters, and has been covered in the popular press by the Wall Street Journal, The Economist, Discover Magazine, New Scientist, Wired, Miller-McCune, the Boston Globe and The Guardian.



Dr. Aaron Clauset

University of Colorado-Boulder

Computer Science *and* Engineering
at the University of Notre Dame
Seminar Series

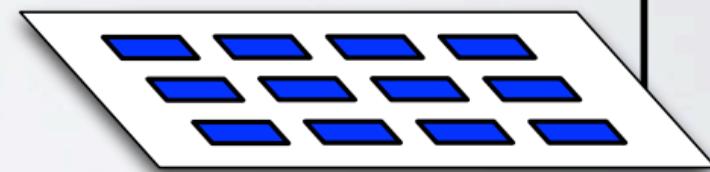
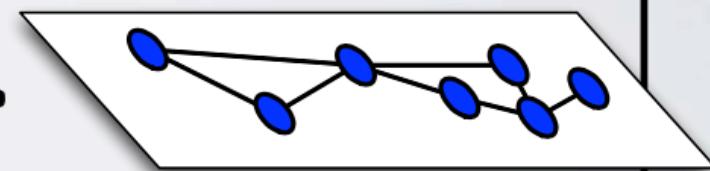




what are networks?

- an approach
- a mathematical representation
- provide structure to complexity
- *structure above*
individuals / components
- *structure below*
system / population

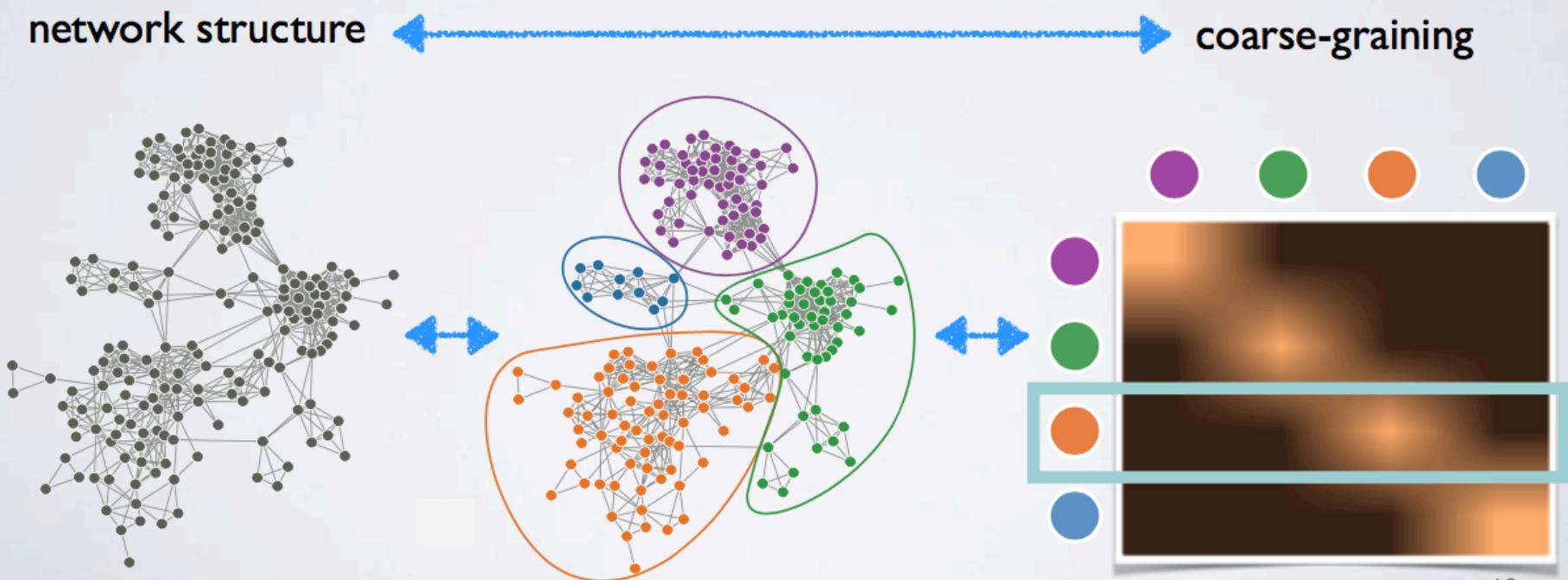
system / population



individuals / components

what is "large-scale structure"?

- coarse-graining of topology
- network modules or communities = building blocks

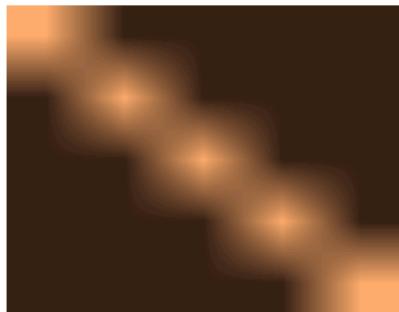


what is "large-scale structure"?

- coarse-graining of topology

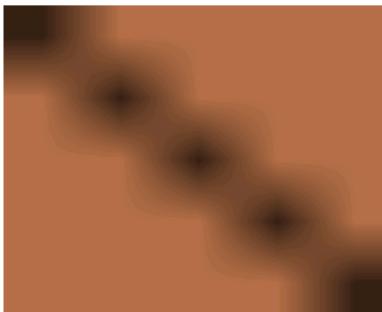
assortative

edges within groups



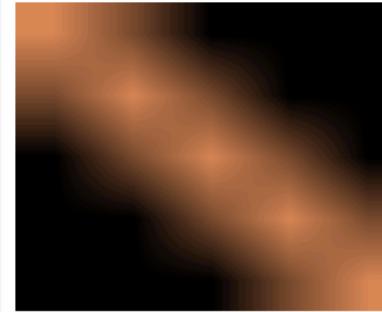
disassortative

edges between groups



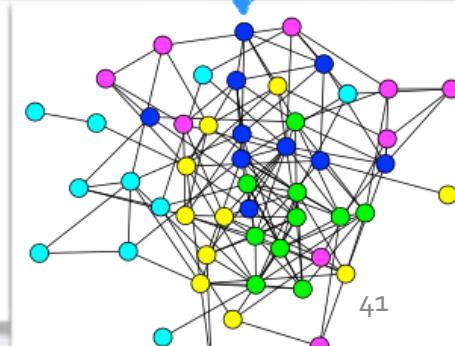
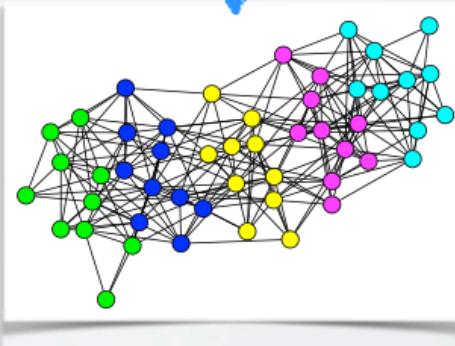
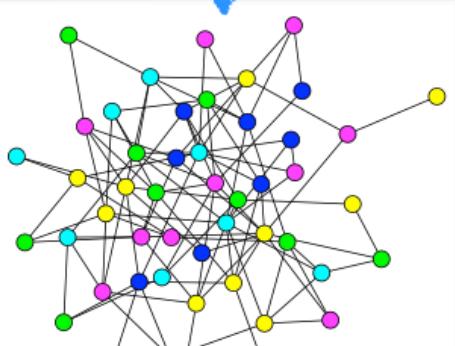
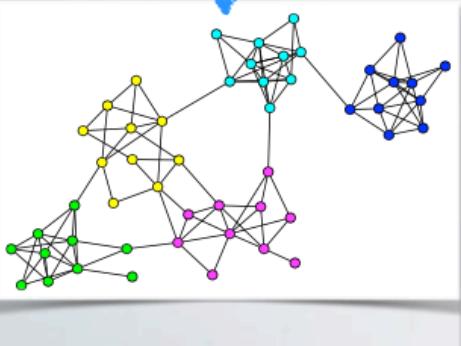
ordered

linear group hierarchy



core-periphery

dense core, sparse periphery





probabilistic generative models for networks

a parametric probability distribution over networks $\Pr(G | \theta)$

general likelihood function :

$$\Pr(G | \theta) = \prod_{ij} \Pr(A_{ij} | \theta)$$

assumptions about “structure” go into $\Pr(A_{ij} | \theta)$

consistency $\lim_{n \rightarrow \infty} \Pr(\hat{\theta} \neq \theta) = 0$



edge generation function

requires that edges be conditionally independent

2

general classes of these models



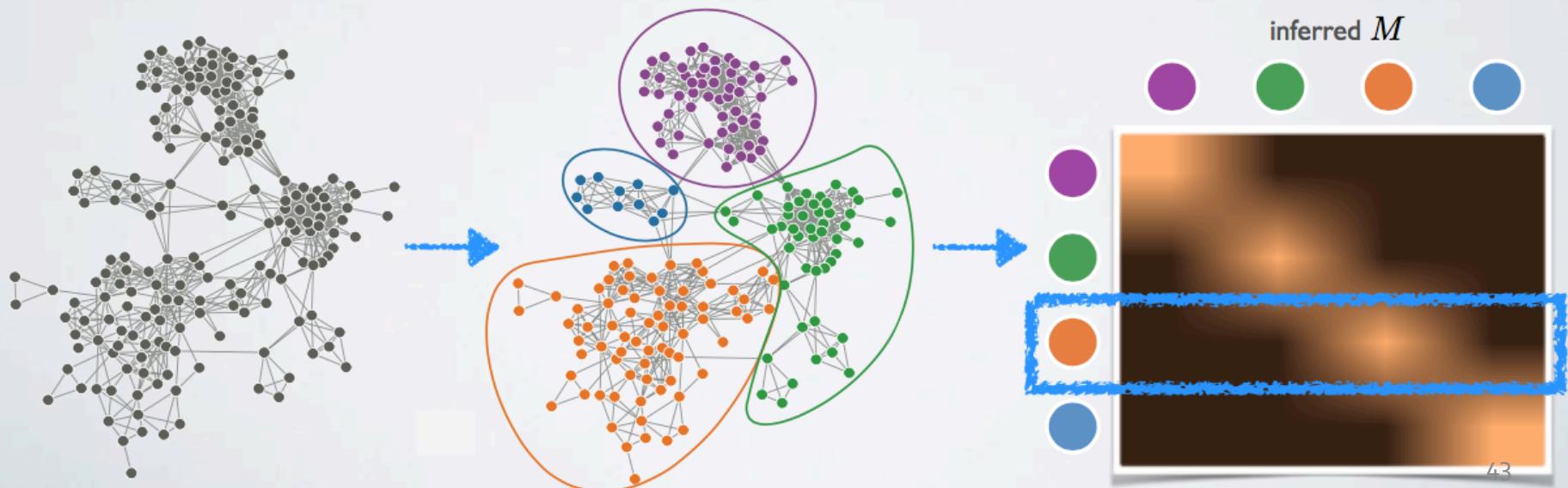
basic stochastic block model

each vertex i has type $z_i \in \{1, \dots, k\}$ (k vertex types or groups)

stochastic block matrix M of group-level connection probabilities

probability that i, j are connected = M_{z_i, z_j}

community = vertices with same pattern of inter-community connections



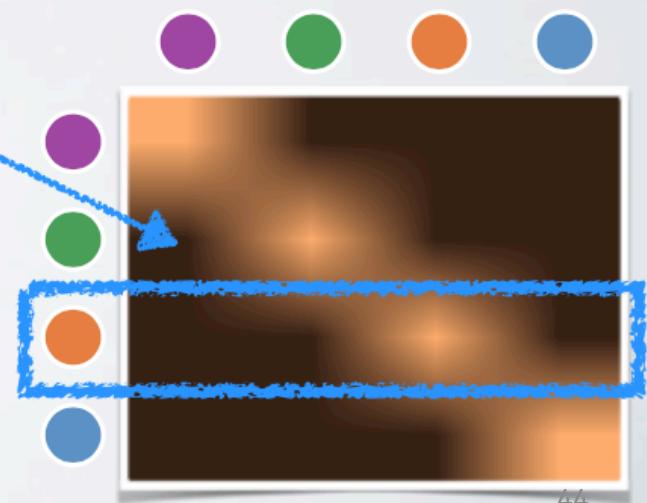


basic stochastic block model

likelihood function :

$$\begin{aligned}\Pr(G \mid z, M) &= \prod_{(i,j) \in E} M_{z_i, z_j} \prod_{(i,j) \notin E} (1 - M_{z_i, z_j}) \\ &= \prod_{rs} M_{r,s}^{e_{r,s}} (1 - M_{r,s})^{n_s n_r - e_{r,s}} \quad (\text{Bernoulli edges})\end{aligned}$$

Bernoulli random graph
with parameter $M_{r,s}$





general stochastic block models

likelihood function :

$$\Pr(A | z, \theta, x) = \prod_{ij} f(A_{ij} | \theta_{\mathcal{R}(z_i, z_j)}, x_i, x_j)$$

A_{ij} : value of adjacency

\mathcal{R} : partition of adjacencies

f : probability function

$\theta_{a,*}$: pattern for a -type adjacencies

x : node attributes (metadata)

Binomial = simple graphs
Poisson = multi-graphs
Normal = weighted graphs
etc.

θ ₁₁	θ ₁₂	θ ₁₃	θ ₁₄
θ ₂₁	θ ₂₂	θ ₂₃	θ ₂₄
θ ₃₁	θ ₃₂	θ ₃₃	θ ₃₄
θ ₄₁	θ ₄₂	θ ₄₃	θ ₄₄



general stochastic block models

naturally models many large-scale patterns

assortative, disassortative, ordered, core-periphery, and various mixtures of these

highly effective in practice

used in gene regulatory networks, protein networks, social networks, food webs, etc.

many nice mathematical features

- general definition of "community" or group
- learns from noisy or missing data
- naturally quantifies uncertainty
- can compare models of large-scale structure [this pattern or that pattern?]
- can predict missing or spurious or future data [link or attribute prediction]
- easily augmentable with auxiliary data
- inferred block matrix is interpretable for science

Typical Clustering Methodologies

- Distance-based methods
- Density-based and grid-based methods
- Probabilistic and generative models
- Leveraging dimensionality reduction methods
- High-dimensional clustering
- Scalable techniques for cluster analysis

User Insights and Interactions in Clustering

- **Visual insights:** One picture is worth a thousand words
 - Human eyes: High-speed processor linking with a rich knowledge-base
 - A human can provide intuitive insights; HD-eye: visualizing HD clusters
- **Semi-supervised insights:** Passing user's insights or intention to system
 - User-seeding: A user provides a number of labeled examples, approximately representing categories of interest
- **Multi-view and ensemble-based insights**
 - Multi-view clustering: Multiple clusterings represent different perspectives
 - Multiple clustering results can be ensembled to provide a more robust solution
- **Validation-based insights:** Evaluation of the quality of clusters generated
 - May use case studies, specific measures, or pre-existing labels

References

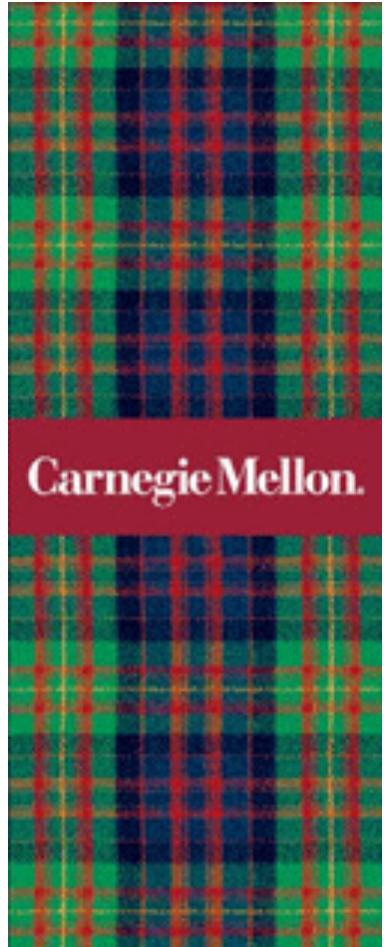
- Jiawei Han, Micheline Kamber, and Jian Pei. Data Mining: Concepts and Techniques. Morgan Kaufmann, 3rd ed. , 2011 (Chapters 10 & 11)
- Charu Aggarwal and Chandran K. Reddy (eds.). Data Clustering: Algorithms and Applications. CRC Press, 2014
- Mohammed J. Zaki and Wagner Meira, Jr.. Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press, 2014
- L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, John Wiley & Sons, 1990
- Charu Aggarwal. An Introduction to Clustering Analysis. in Aggarwal and Reddy (eds.). Data Clustering: Algorithms and Applications (Chapter 1). CRC Press, 2014

CATCHTARTAN: Representing and Summarizing Dynamic Multicontextual Behaviors

Meng Jiang, Christos Faloutsos, Jiawei Han



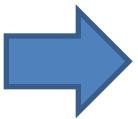
What is Tartan?



GO TARTANS!



Visited CMU in 2012-13



Watched lots of
Tartans' games...



What is Behavior? Is it valuable?

Behavior: interactions made by **individuals or organisms** in conjunction with **themselves** or their **environment**. (Wikipedia)

- Tweeting behavior
- Publishing-paper behavior

20:03:09 @ebekahwsm
this better be the best halftime show ever in the history of halftimes shows. ever.
#SuperBowl

2009 P. Melville, W. Gryc, R. Lawrence,
“Sentiment analysis of blogs by combining lexical knowledge with text classification”, KDD’09. Refs: p81623, p84395...

Q: What can we discover from behavioral data?

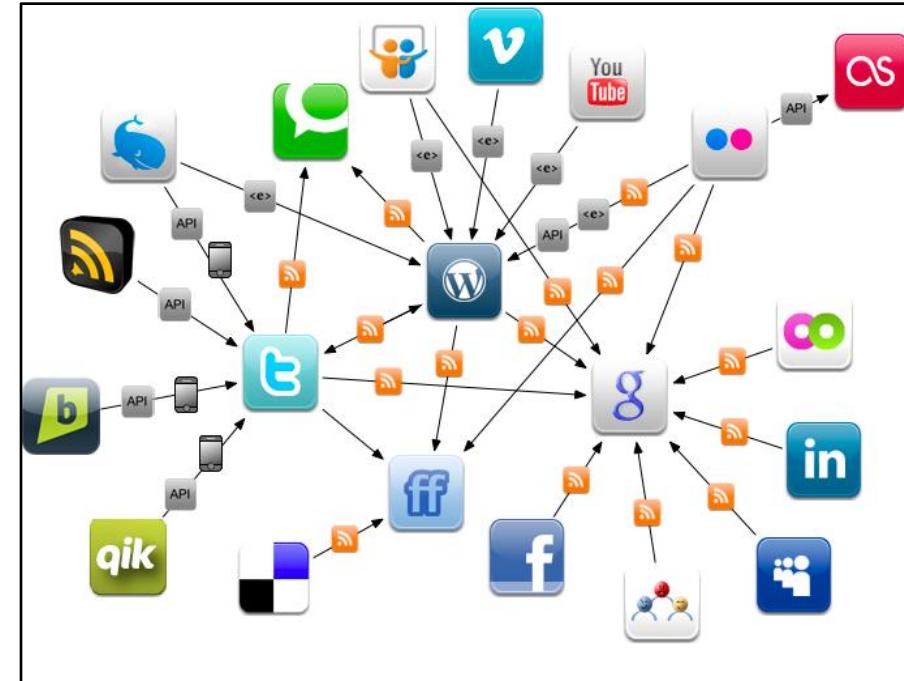
Ex. Given every phone call / message between the military leaders, scientists, businesspersons, Find ...

Why We Talk about Behavior Today?

Physical Environment



Online Environment



The human behaviors are broadly and deeply recorded in an unprecedented level.

Representing and Summarizing Behavior

Representing

Raw data to Math

Summarizing

Patterns: trends, events, campaigns...

Understanding

Factors underlying the patterns: influence, intentions...

Predicting

What will happen in the future?

Intervening

Recommendation, spam/fraud detection...

Given the behavioral data (e.g., DBLP data, tweets)

2009 P. Melville, W. Gryc, R. Lawrence, “Sentiment analysis of blogs by combining lexical knowledge with text classification”, KDD’09. Refs: p81623, p84395...

Return behavioral summaries (e.g., research trends, events)

Author	Venue	Keyword	Cited	#Paper	Venue	Keyword	#Paper
76 Cheng-xiang Zhai Hui Fang S. Kambhampati	7 SIGIR VLDB TKDE	7 “information retrieval” “data integration” “text classification”	68 p56743 ¹ p62995 p76869	32 2003- 2007	5 ICML NIPS ...	6 “reinforcement learning” “machine learning”	40 1997- 2002

¹ “A language modeling approach to information retrieval”

Author	Venue	Cited	#Paper	Venue	Keyword	#Paper	Author	Venue	Keyword	#Paper
6 Jiawei Han Xifeng Yan	1 SIG-MOD	1 p76095 ²	22 2004- 2010	3 ICDM AAAI TKDE	1 “anomaly detection”	25 2005- 2013	27 C. Faloutsos J. Pei P. S. Yu X. Lin C. Aggarwal...	6 KDD ICDM ICDE TKDE ...	12 “large graphs” “data streams” “evolving data” “evolving graphs” ...	70 2006- 2013

² “Frequent subgraph discovery”

Author	Venue	Keyword	Cited	#Paper	Author	Venue	Keyword	#Paper
12 Ryen White Hang Li Tie-Yan Liu Zhaohui Zheng...	5 SIGIR WWW WSDM CIKM...	3 “web search” “click-through data” “sponsored search”	12 p82630 ³ p116290 p103899 p106191...	32 2006- 2013	8 Qiang Yang Dou Shen Sinno Pan...	3 KDD PAKDD AAAI	6 “transfer learning” “data mining” “localization models”	17 2007- 2010

³ “Optimizing search engines using clickthrough data”

Behaviors: Dynamic and Multicontextual

- Tweeting behavior

20:03:09 @ebekahwsm
this better be the best halftime show ever
in the history of halftimes shows. ever.
#SuperBowl

Contextual factors:

*One-guaranteed
value*

value



*Empty (set
of) value*



Set value



*Empty (set
of) value*



Dynamic

Time slice	User	Location	Phrase	Hashtag	URL
20:00- 20:30	@ebekahwsm	∅	{best halftime show, in the history, halftimes shows}	{#SuperBowl}	∅

Behaviors: Dynamic and Multicontextual

- Publishing-paper behavior

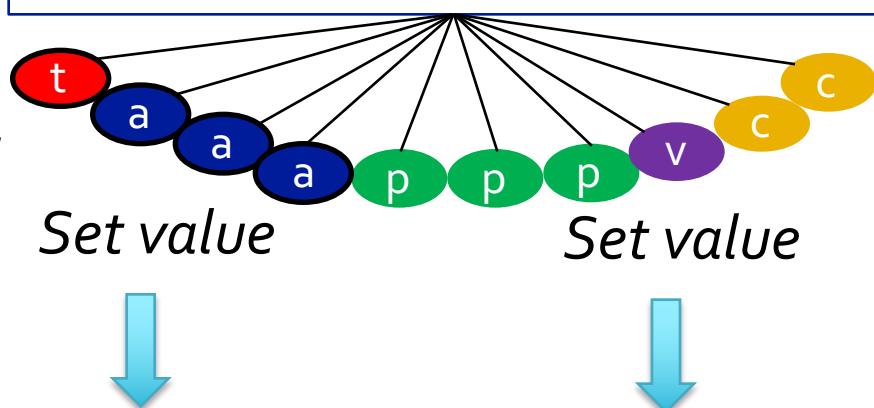
2009 P. Melville, W. Gryc, R. Lawrence,
“Sentiment analysis of blogs by combining lexical knowledge with text classification”, KDD’09. Refs: p81623, p84395...

Contextual factors:

*One-guaranteed
value*

Set value

Dynamic



Time slice	Author	Venue	Keyword	Cited papers
2009	{P. Melville, W. Gryc, R. Lawrence}	SIGKDD	{sentiment analysis, lexical knowledge, text classification}	{p81623, p84395, p95393, p95409, p99073, p116349 ...}

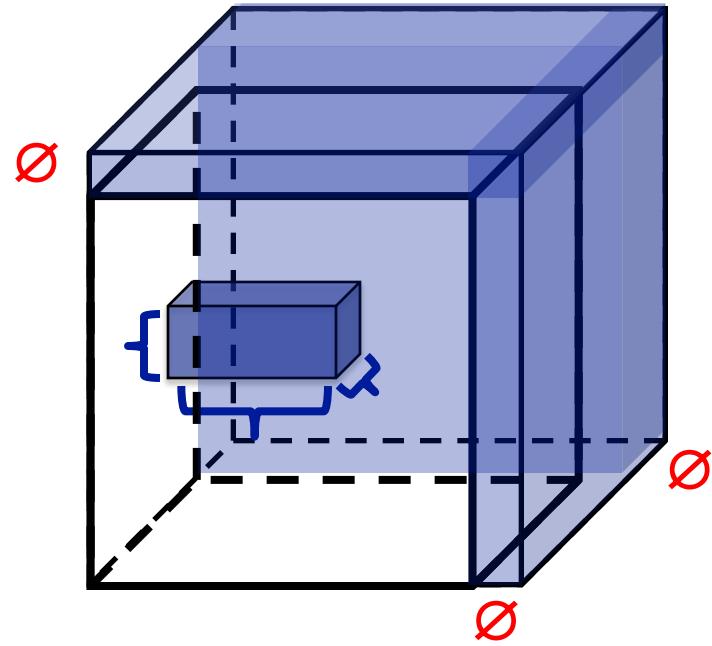
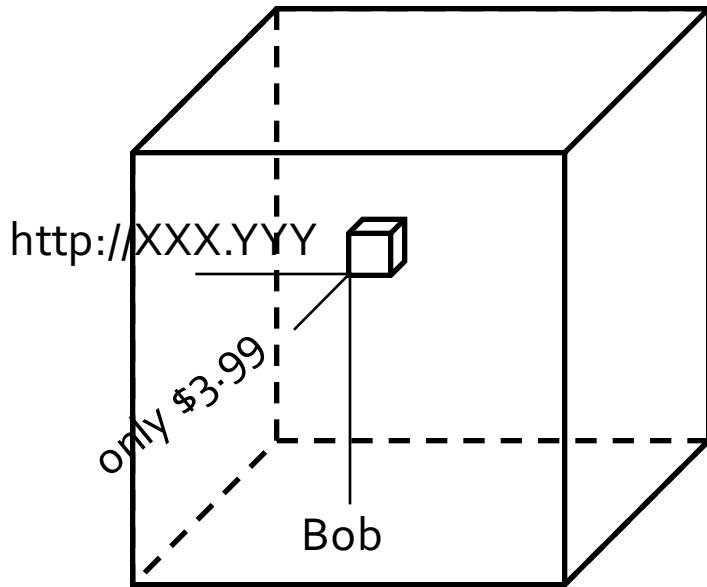
Summarizing Behaviors

- *Dynamic*: taking a set of consecutive time slices
- *Multicontextual*: taking a set of dimensions and a set of dimensional values in each dimension

Term	Definition
Dimension	The type of a contextual factor (e.g., location, phrase; author, keyword)
(Dimensional) value	The contextual factor in the dimension
Time slice	The period for consecutive behaviors
Behavior	A set of dimensions, a set of values in each dimension, a time slice for the timestamp

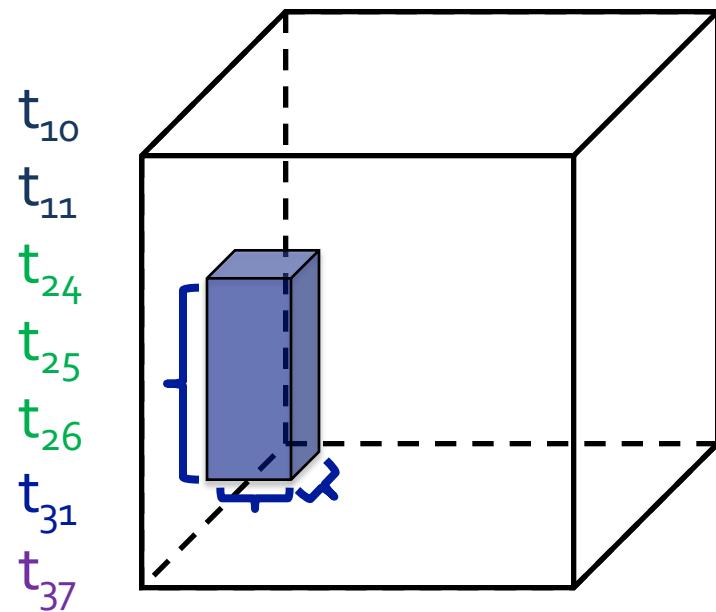
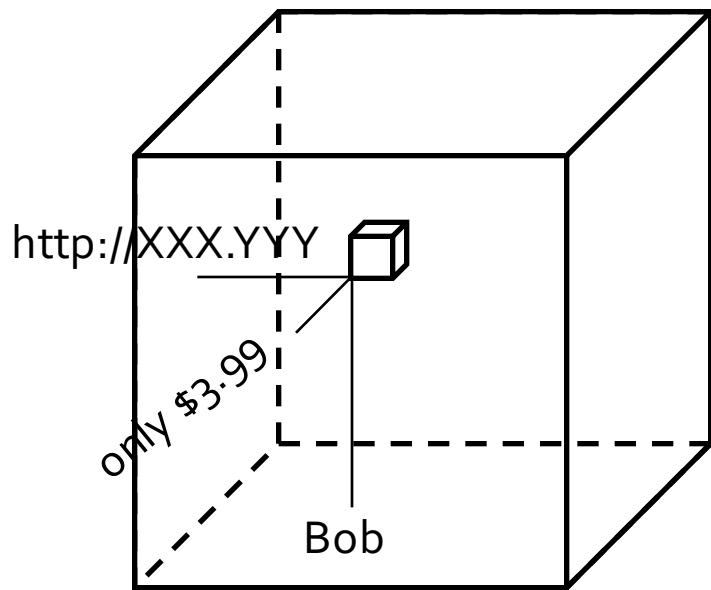
Tensor Fails

- Tensor - modeling multidimensions: FEMA (KDD'14), CrossSpot (ICDM'15)
- Representation: (multicontextual)
 - Empty values?



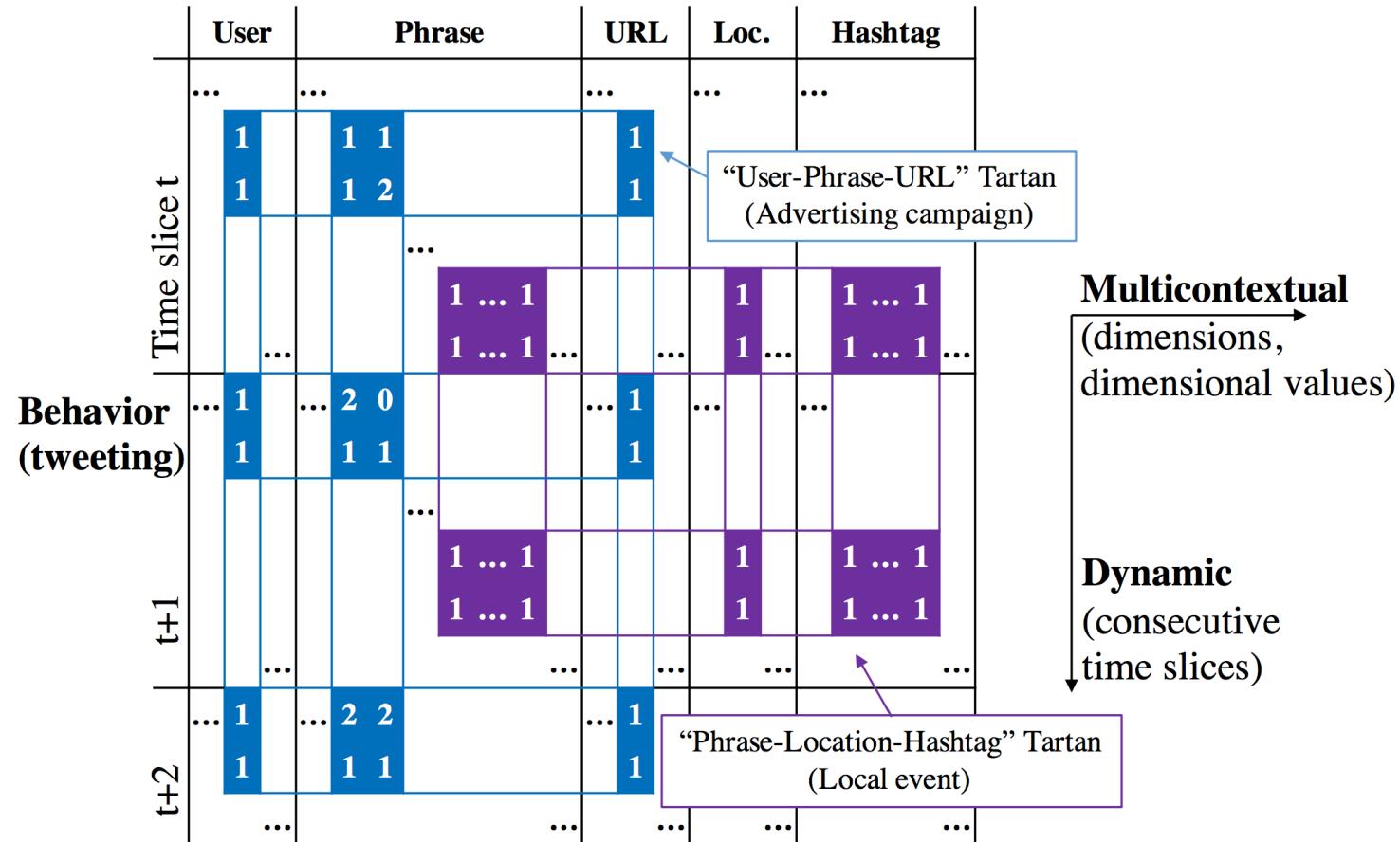
Tensor Fails (cont.)

- Tensor - modeling multidimensions: FEMA (KDD'14), CrossSpot (ICDM'15)
- **Summarization: (dynamic)**
 - Temporal patterns?



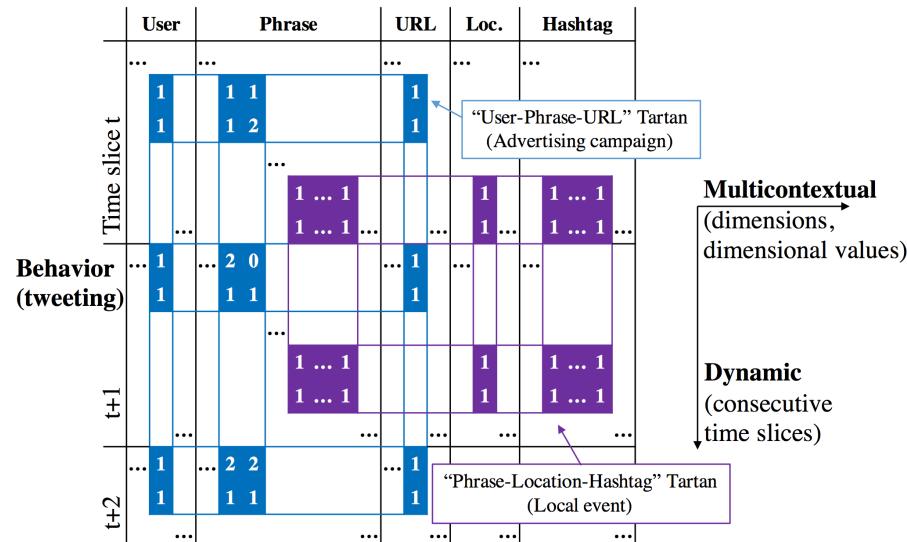
Our Representations for Behavior and Behavioral Summary

- Behavior: “Two-level matrix”
- Behavioral summary: “Tartan”



The Problem of Behavioral Summarization

PROBLEM 1 (BEHAVIORAL SUMMARIZATION). *Given the behavioral data (a two-level matrix) $\mathcal{X} = \{D, N_d|_{d=1}^D, T, E^{(t)}|_{t=1}^T\}$, find a list of behavioral summaries (Tartans) $\tilde{\mathcal{A}} = \{\dots, \mathcal{A}, \dots\}$ ordered by a principled metric function $f(\mathcal{A}, \mathcal{X})$ which defines how well the sets of meaningful dimensions, values, time slices and behaviors are partitioned and how well the meaningful subset of data is summarized, where $\mathcal{A} = \{\mathcal{D}, \mathcal{V}_d|_{d \in \mathcal{D}}, \mathcal{T}, \mathcal{B}^{(t)}|_{t \in \mathcal{T}}\}$.*



CATCHTARTAN

- Employing a lossless encoding scheme
 - The *Minimum Description Length* (MDL) principle
 - Estimating the **number of bits** that encoding the Tartan can **save from** merging the meaningful pattern into the encoding of the data

	FSG [18]	GRAPH- CUBE [33]	EVENT- CUBE [29]	MDC [21]	BoW [6]	FEMA [9]	COM2 [2]	CROSS- SPOT [8]	GRAPH- SCOPE [27]	VoG [15]	TIME- CRUNCH [26]	CATCH- TARTAN
Principled scoring	✓							✓	✓	✓	✓	✓
Parameter-free		✓						✓	✓	✓	✓	✓
Multidimensional			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Multicontextual				✓	✓							✓
Timestamp value						✓		✓			✓	✓
Dynamics							✓	✓	✓	✓	✓	✓

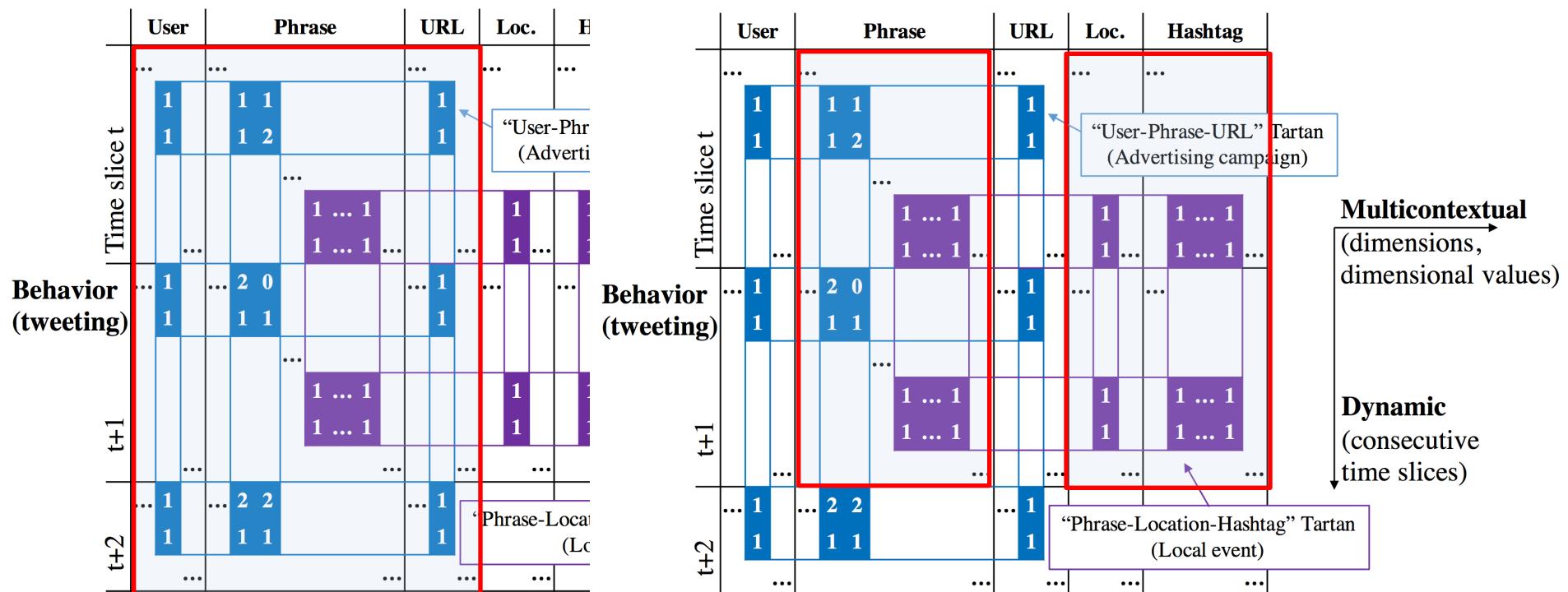
Objective Function to Maximize

$$f(\mathcal{A}, \mathcal{X}) = L(\mathcal{X}^{\mathcal{A}}) - L(\mathcal{A}) - L(\mathcal{X}^{\mathcal{A}} \setminus \mathcal{A}).$$

Tartan

Data First-level matrix

Individual entries



$$\mathcal{X}^{\mathcal{A}} = \{\mathcal{X}_d^{(t)}(b, i) | d \in \mathcal{D}, t \in \mathcal{T}, i \in \{1, \dots, N_d\}, b \in \{1, \dots, E^{(t)}\}\}.$$

Objective Function to Maximize (cont.)

$$f(\mathcal{A}, \mathcal{X}) = L(\mathcal{X}^{\mathcal{A}}) - L(\mathcal{A}) - L(\mathcal{X}^{\mathcal{A}} \setminus \mathcal{A}).$$

$$V = (\sum_{d \in \mathcal{D}} N_d) (\sum_{t \in \mathcal{T}} E^{(t)}).$$

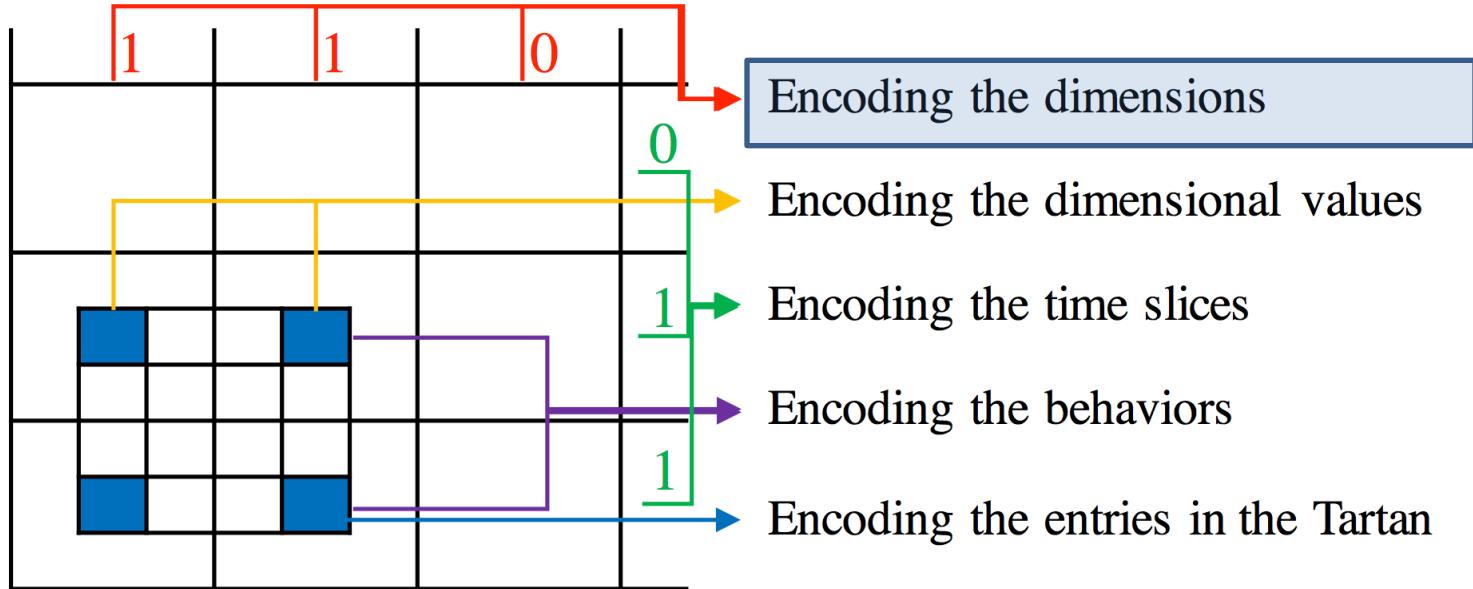
$$C = \sum_{d \in \mathcal{D}, t \in \mathcal{T}} \sum_{b \in \{1, \dots, E^{(t)}\}, i \in \{1, \dots, N_d\}} \mathcal{X}_d^{(t)}(b, i).$$

$$\begin{aligned} L(\mathcal{X}^{\mathcal{A}}) &= g(V + C, C) + L_{\mathcal{D}}(\mathcal{A}) + L_{\mathcal{T}}(\mathcal{A}) \\ &\quad + \sum_{d \in \mathcal{D}} \log^* N_d + \sum_{t \in \mathcal{T}} \log^* E^{(t)}. \end{aligned}$$

$$L(\mathcal{A}) = L_{\mathcal{D}}(\mathcal{A}) + L_{\mathcal{V}}(\mathcal{A}) + L_{\mathcal{T}}(\mathcal{A}) + L_{\mathcal{B}}(\mathcal{A}) + L_{\mathcal{A}}(\mathcal{A}).$$

$$L(\mathcal{X}^{\mathcal{A}} \setminus \mathcal{A}) = g(V + C - v - c, C - c);$$

Encoding the Tartan: Dimensions



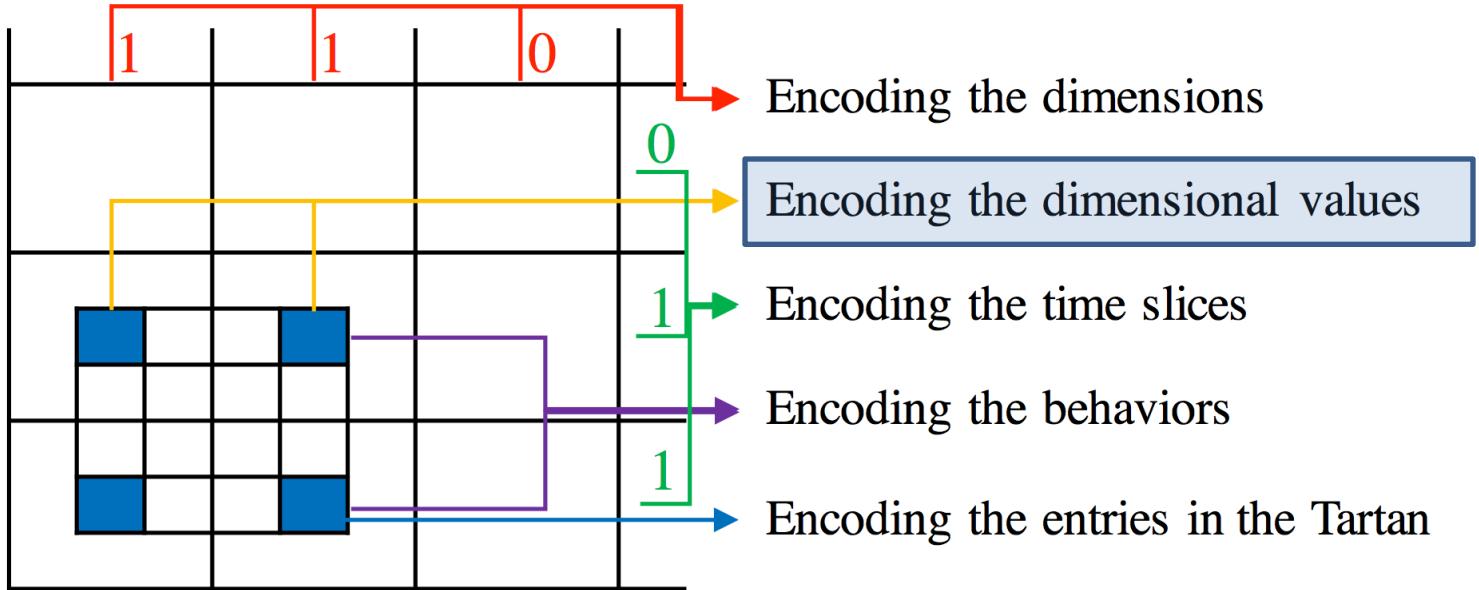
$$H_{\mathcal{D}}(X) = - \sum_{x \in \{0,1\}} P(X = x) \log P(X = x)$$

$$= - \left(\frac{D^{\mathcal{A}}}{D} \log \frac{D^{\mathcal{A}}}{D} + \frac{D-D^{\mathcal{A}}}{D} \log \frac{D-D^{\mathcal{A}}}{D} \right).$$

$$L_{\mathcal{D}}(\mathcal{A}) = \log^* D + \log^* D^{\mathcal{A}} + D \cdot H_{\mathcal{D}}(X)$$

$$= \log^* D + \log^* D^{\mathcal{A}} + g(D, D^{\mathcal{A}}),$$

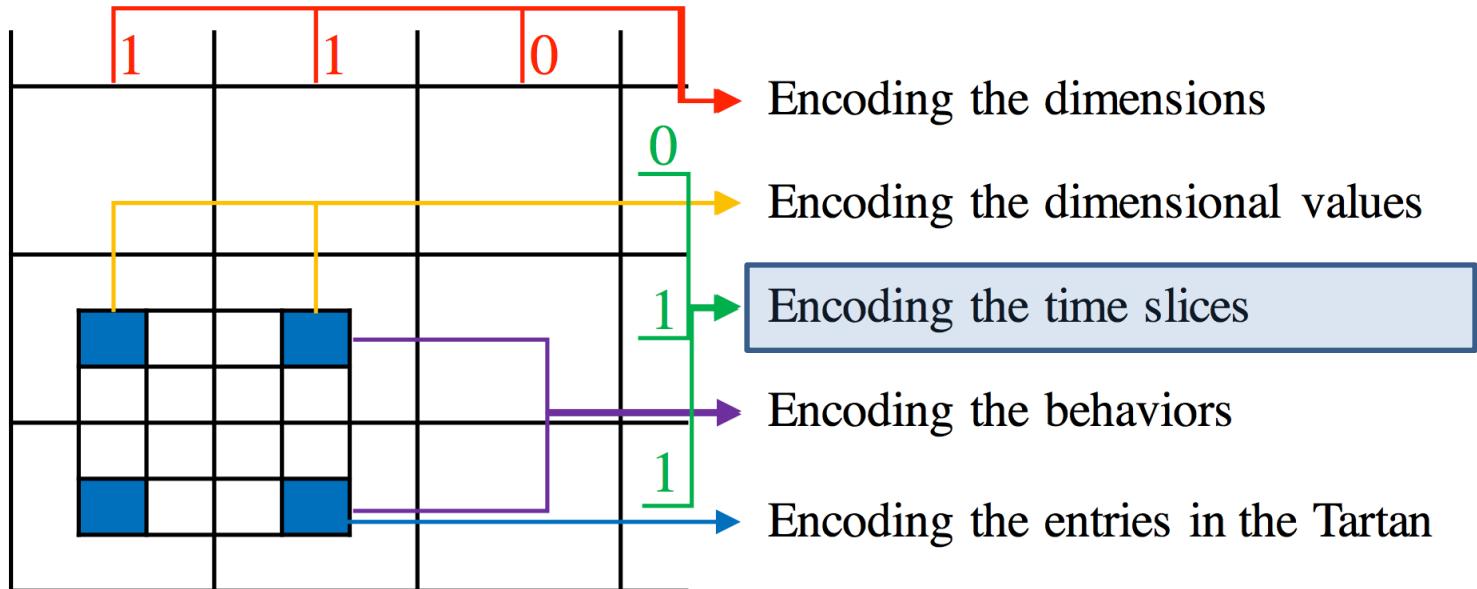
Encoding the Tartan: Dimensional Values



$$H_{\mathcal{V}_d}(X) = - \left(\frac{n_d}{N_d} \log \frac{n_d}{N_d} + \frac{N_d - n_d}{N_d} \log \frac{N_d - n_d}{N_d} \right).$$

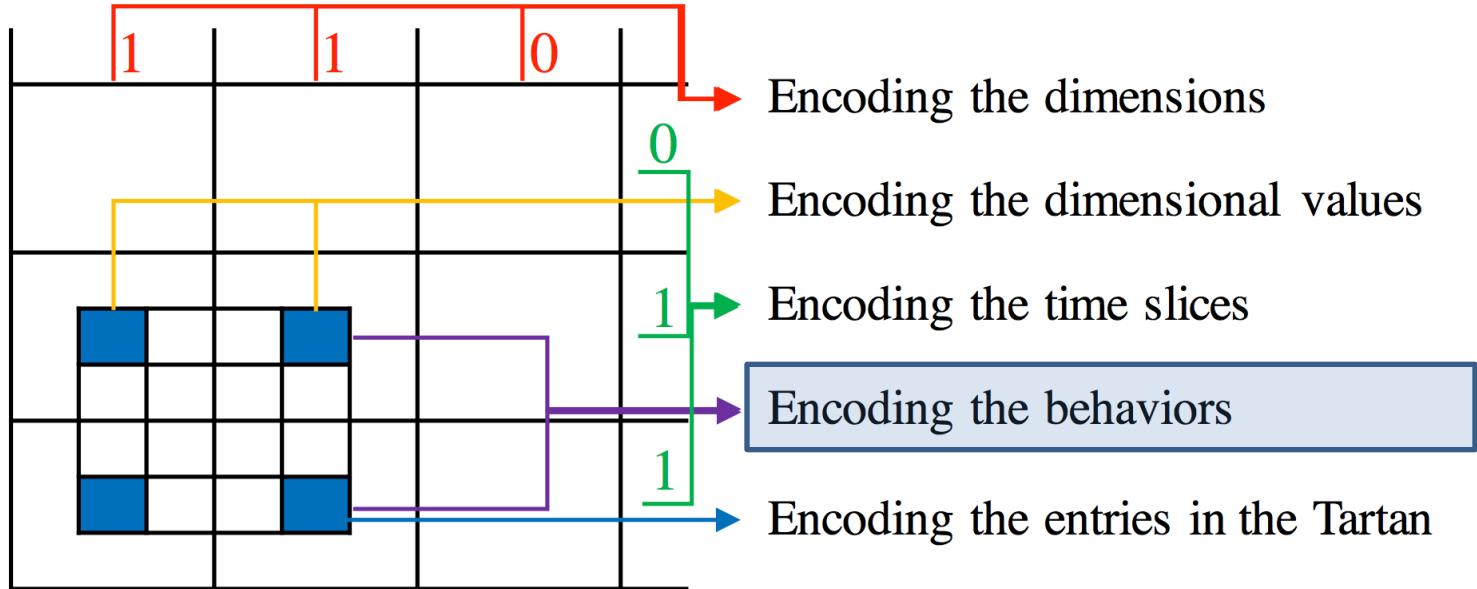
$$L_{\mathcal{V}}(\mathcal{A}) = \sum_{d \in \mathcal{D}} \left(\log^* N_d + \log^* n_d + g(N_d, n_d) \right).$$

Encoding the Tartan: Time Slices



$$L_{\mathcal{T}}(\mathcal{A}) = \log^* T + \log^* T^{\mathcal{A}} + \log^* t_{start}$$

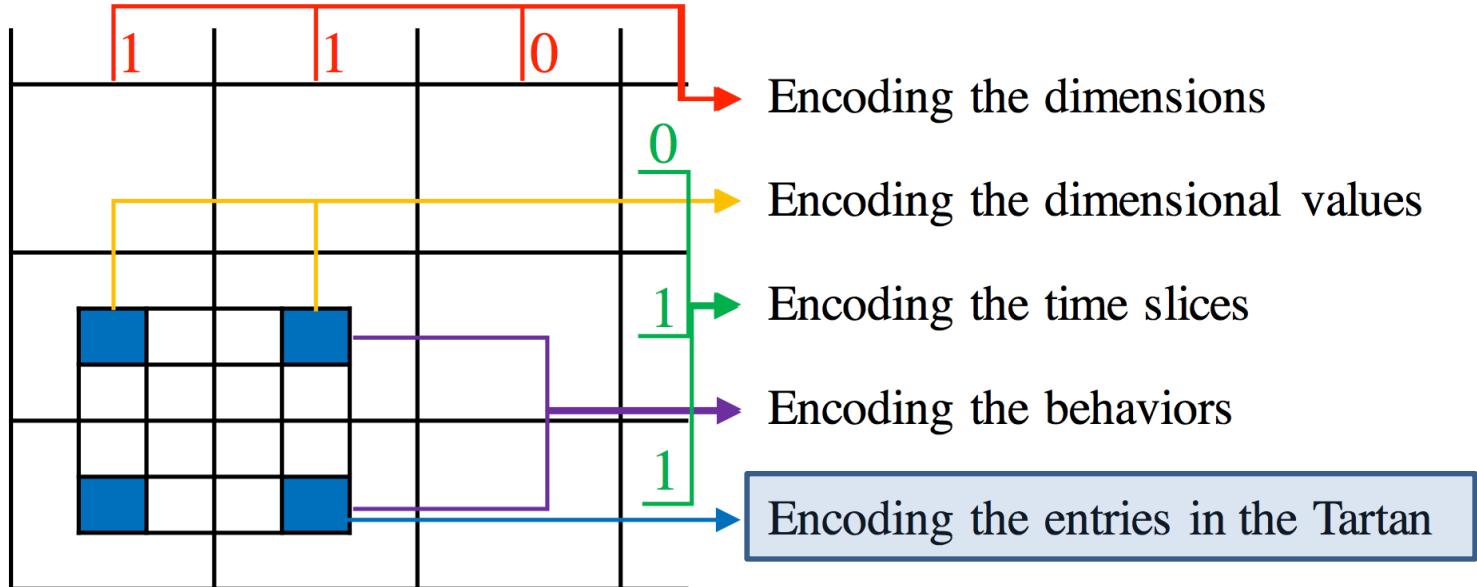
Encoding the Tartan: Behaviors



$$H_{\mathcal{B}^{(t)}}(X) = - \left(\frac{e^{(t)}}{E^{(t)}} \log \frac{e^{(t)}}{E^{(t)}} + \frac{E^{(t)} - e^{(t)}}{E^{(t)}} \log \frac{E^{(t)} - e^{(t)}}{E^{(t)}} \right).$$

$$L_{\mathcal{B}}(\mathcal{A}) = \sum_{t \in \mathcal{T}} \left(\log^* E^{(t)} + \log^* e^{(t)} + g(E^{(t)}, e^{(t)}) \right).$$

Encoding the Tartan: Entries



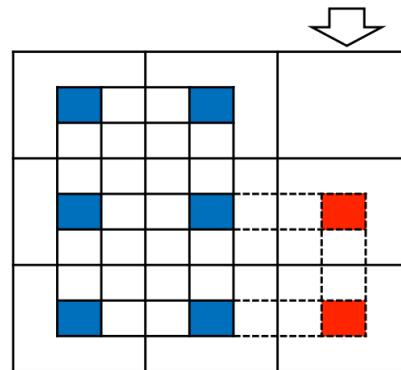
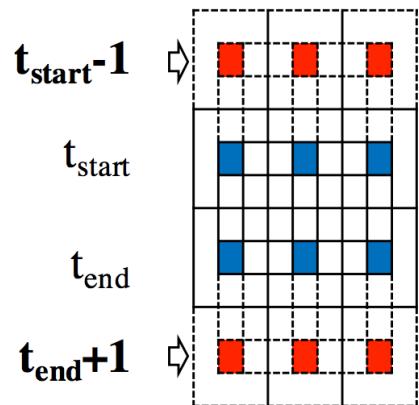
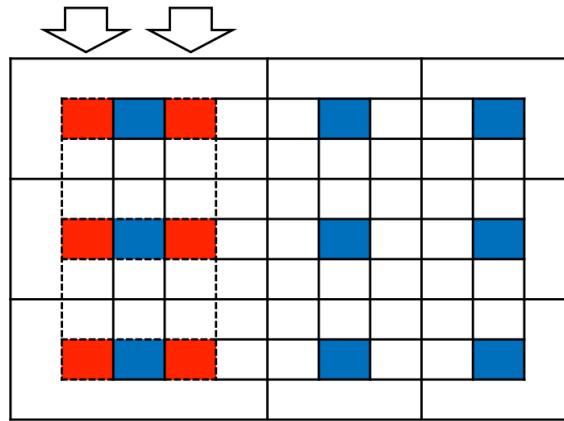
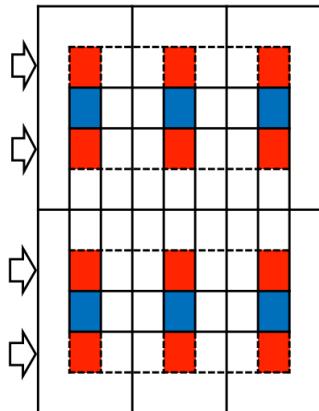
$$v = \left(\sum_{d \in \mathcal{D}} n_d \right) \left(\sum_{t \in \mathcal{T}} e^{(t)} \right).$$

$$c = \sum_{d \in \mathcal{D}, t \in \mathcal{T}} \sum_{b \in \mathcal{B}^{(t)}, i \in \mathcal{V}_d} \chi_d^{(t)}(b, i).$$

$$H_{\mathcal{A}}(X) = - \left(\frac{c}{v+c} \log \frac{c}{v+c} + \frac{v}{v+c} \log \frac{v}{v+c} \right).$$

$$L_{\mathcal{A}}(\mathcal{A}) = (v + c) H_{\mathcal{A}}(X) = g(v + c, c).$$

Greedy Search for the Local Minimum



Time complexity:

$$\mathcal{O}(\sum_d N_d \log N_d + \sum_t E^{(t)} \log E^{(t)})$$

Qualitative Analysis: DBLP data

Author	Venue	Keyword	Cited	#Paper
76 Cheng-xiang Zhai Hui Fang S. Kambhampati	7 SIGIR VLDB TKDE	7 “information retrieval” “data integration” “text classification”	68 p56743 ¹ p62995 p76869	32 2003-2007

Venue	Keyword	#Paper
5 ICML NIPS ...	6 “reinforcement learning” “machine learning”	40 1997-2002

¹ “A language modeling approach to information retrieval”

Author	Venue	Cited	#Paper	Venue	Keyword	#Paper
6 Jiawei Han Xifeng Yan	1 SIG-MOD	1 p76095 ²	22 2004-2010	3 ICDM AAAI TKDE	1 “anomaly detection”	25 2005-2013

² “Frequent subgraph discovery”

Author	Venue	Keyword	#Paper
27 C. Faloutsos J. Pei P. S. Yu X. Lin C. Aggarwal...	6 KDD ICDM ICDE TKDE ...	12 “large graphs” “data streams” “evolving data” “evolving graphs” ...	70 2006-2013

Author	Venue	Keyword	Cited	#Paper
12 Ryen White Hang Li Tie-Yan Liu Zhaohui Zheng...	5 SIGIR WWW WSDM CIKM...	3 “web search” “click-through data” “sponsored search”	12 p82630 ³ p116290 p103899 p106191...	32 2006-2013

³ “Optimizing search engines using clickthrough data”

Author	Venue	Keyword	#Paper
8 Qiang Yang Dou Shen Sinno Pan...	3 KDD PAKDD AAAI	6 “transfer learning” “data mining” “localization models”	17 2007-2010

1997 2000 2003 2006 2009 2012

Qualitative Analysis: Super Bowl 2013

16:30	16:30:31 <u>My prediction</u> Ravens 34 Niners 31 16:30:57 Ready for the big game :D, <u>my prediction</u> 24-20 SF #SuperBowl 16:31:14 <u>My prediction for superbowl..</u> 48.. Jets over Bears 17-13 Mark Sanchez MVP 16:32:24 I predict <u>Baltimore Ravens</u> will win 27 to 24 or 25 or 26. Basically it will be a close game.	“my prediction”	user	phrase	hashtag	URL	3,397 tweets	Tartan #1: (1 dim) 16:30-17:30	
17:00			(3,325)	226	(0)	(0)			
17:30	17:30:51 RT @LMAOTWITPICTS: <u>Make Your Prediction</u> . Retweet For 49ers http://t.co/KKksEist 17:31:01 RT @LMAOTWITPICTS: <u>Make Your Prediction</u> . Retweet For 49ers http://t.co/KKksEist 17:31:16 RT @LMAOTWITPICTS: <u>Make Your Prediction</u> . Retweet For 49ers http://t.co/KKksEist 17:31:19 RT @LMAOTWITPICTS: <u>Make Your Prediction</u> . Retweet For 49ers http://t.co/KKksEist	“make your prediction”	user	phrase	RT @user	URL	196 tweets	Tartan #2: (3 dims) 17:00-18:00	
18:00	18:55:03 RT @49ers: Kaepernick is sacked on 3rd and goal. #49ers K David Akers makes 36-yard FG. Baltimore leads 7-3 with 3:58 left in 1st Qtr. #SB47 18:55:04 RT @49ers: Kaepernick is sacked on 3rd and goal. #49ers K David Akers makes 36-yard FG. Baltimore leads 7-3 with 3:58 left in 1st Qtr. #SB47 18:55:44 RT @Ravens: David Akers is good from 36 yards to make the score 7-3 Ravens. Nice job by the defense to tighten up in the red zone.	“7-3”, “1 st Qtr”	user	phrase	RT @user	URL	215 tweets	Tartan #3: (2 dims) 18:30-19:30	
18:30			(213)	21	3	(0)			
19:00									
19:30	20:20:01 RT @ExtraGrumpyCat: No Superbowl halftime show will ever surpass this. http://t.co/0VSy7Cv6 20:20:02 RT @WolfpackAlan: No Superbowl halftime show will ever surpass this. http://t.co/6BlloPXs 20:20:04 RT @ExtraGrumpyCat: No Superbowl halftime show will ever surpass this. http://t.co/0VSy7Cv6 20:20:05 RT @WolfpackAlan: No Superbowl halftime show will ever surpass this. http://t.co/6BlloPXs	halftime show”	user	phrase	RT @user	URL	617 tweets	Tartan #4: (3 dims) 20:00-21:00	
20:00	20:20:47 (Manhattan, NY)...and every one of those girls took #ballet #Beyonce #superbowl 20:22:01 (New York, NY) I have the biggest lady boner for Beyonce #BeyonceBowl #DestinyBowl #DestinysChild #SuperBowl		location	phrase	hashtag	URL	166 tweets	Tartan #5: (3 dims) 20:00-21:00	
20:30	20:24:32 (Manhattan, NY) No one can ever top that performance by Beyonce. #superbowl, #DestinysChild EVER. #Beyonce #superbowl #halftimeshow	“beyonce”, #beyonce,	2	55	17	(0)			
21:00	21:44:42 Ahora si pff #49ers 23-28 #Ravens 21:44:44 Baltimore #Ravens 28-23 San Francisco #49ers 21:44:50 FG Akers #49ers 23-28 #Ravens 3Q 3:10 #SuperBowlXLVII #SuperBowl #NFL	“28-23”, #49ers, #Ravens	user	phrase	hashtag	URL	653 tweets	Tartan #6: (2 dims) 21:00-22:00	
21:30			(650)	69	11	(0)			
22:00	22:42:27 Congratulations Ravens!!!! 22:42:43 Congratulations Ray Lewis and the Ravens. 22:42:43 Game over! Ravens won ray got his retirement ring now all y'all boys and girls go to sleep ! 22:42:52 @LetThatBoyTweet: Game over. Ravens win the Super Bowl.”	“congratulations”, “game over”	user	phrase	hashtag	URL	1,950 tweets	Tartan #7: (1 dim) 22:00-23:30	

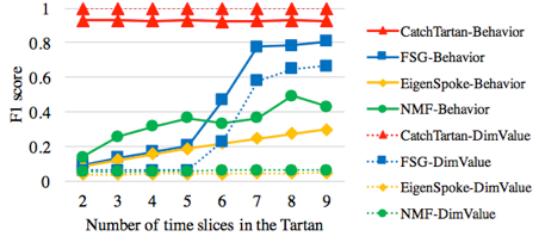
Quantitative Analysis: Accuracy and Efficiency in Synthetic Experiments

- Tartan distribution

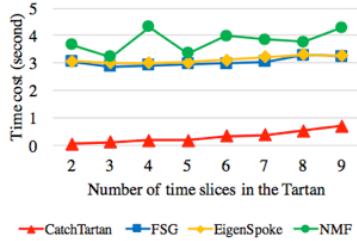
1. $T^{\mathcal{A}} \in [2, 9]$, the number of consecutive time slices in the Tartan \mathcal{A} , 4 as default;
2. $e^{(t)} \in [100, 2,000]$, the number of behaviors in the time slice, 1,000 as default;
3. $D^{\mathcal{A}} \in [2, 9]$, the number of dimensions in \mathcal{A} , 3 as default;
4. $n_d \in [50, 200]$, the number of values per dimension in \mathcal{A} , 100 as default;
5. $\rho \in [1, 10]$, the average number of values per dimension in the behaviors, 3 as default;

- Data distribution

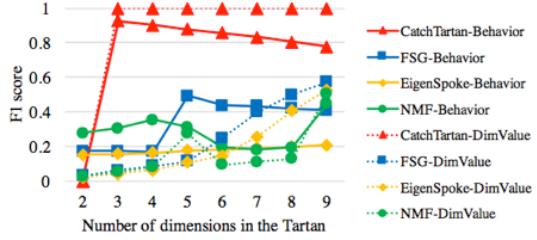
6. $T \in [5, 30]$, the total number of time slices in the dataset, 10 as default;
7. $E^{(t)} \in [1,000, 10,000]$, the number of behaviors per time slice in the dataset, 5,000 as default;
8. $N_d \in [1,000, 2,000]$, the number of values per dimension in the data, 1,000 as default.



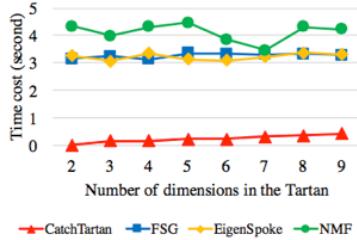
(a) F1 score vs T^A .



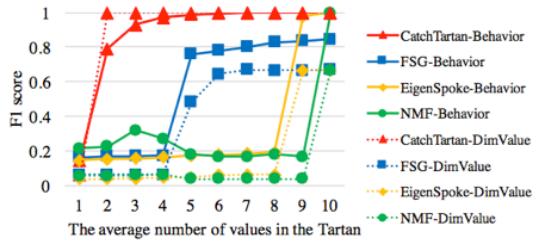
(b) Time cost vs T^A .



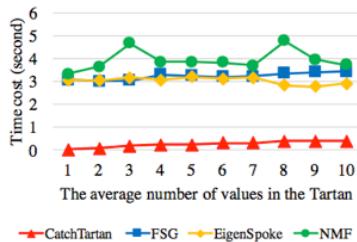
(e) F1 score vs D^A .



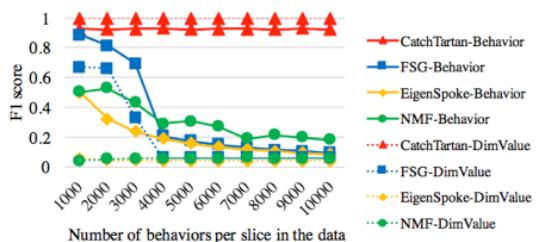
(f) Time cost vs D^A .



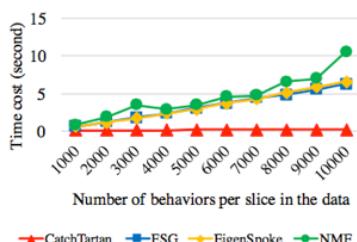
(i) F1 score vs ρ .



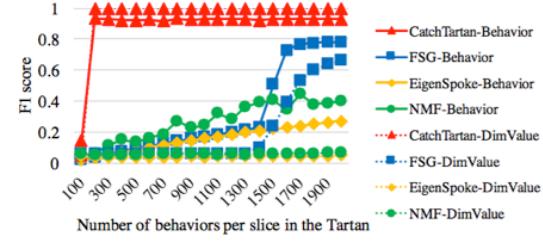
(j) Time cost vs ρ .



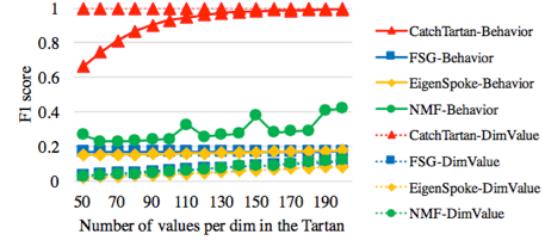
(m) F1 score vs $E^{(t)}$.



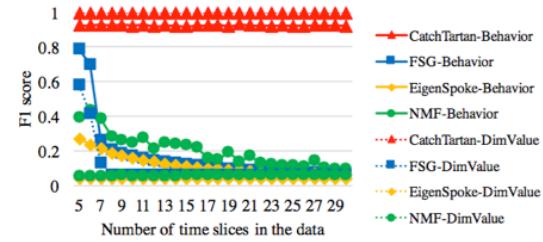
(n) Time cost vs $E^{(t)}$.



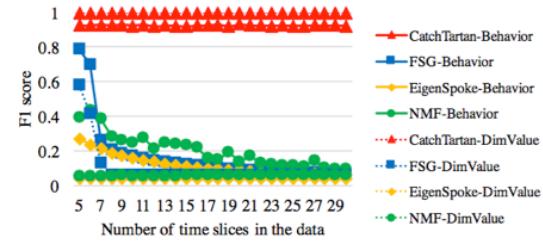
(c) F1 score vs $e^{(t)}$.



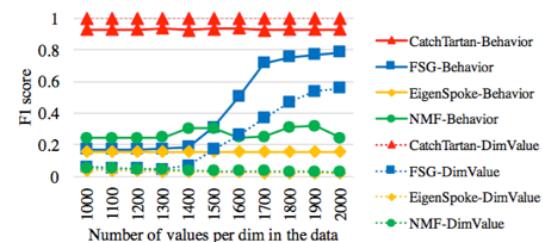
(d) Time cost vs $e^{(t)}$.



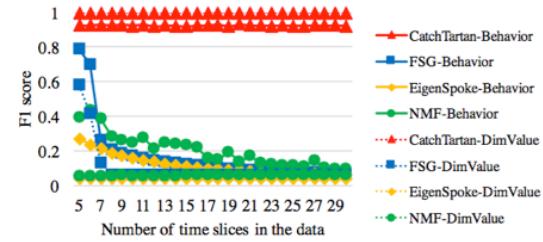
(h) Time cost vs n_d .



(l) Time cost vs T .



(o) F1 score vs N_d .



(p) Time cost vs N_d .

Summary

- Novel representations
 - Behavior: “two-level matrix” vs. tensor
 - Behavioral summary: “Tartan” vs. dense block
- A new summarization algorithm
 - Principled-scoring and Parameter-free: Objective function based on Minimum Description Length
 - Scalable: Greedy search for local optimum
- Effectiveness, discovery and efficiency