BUENOS AIRES 15

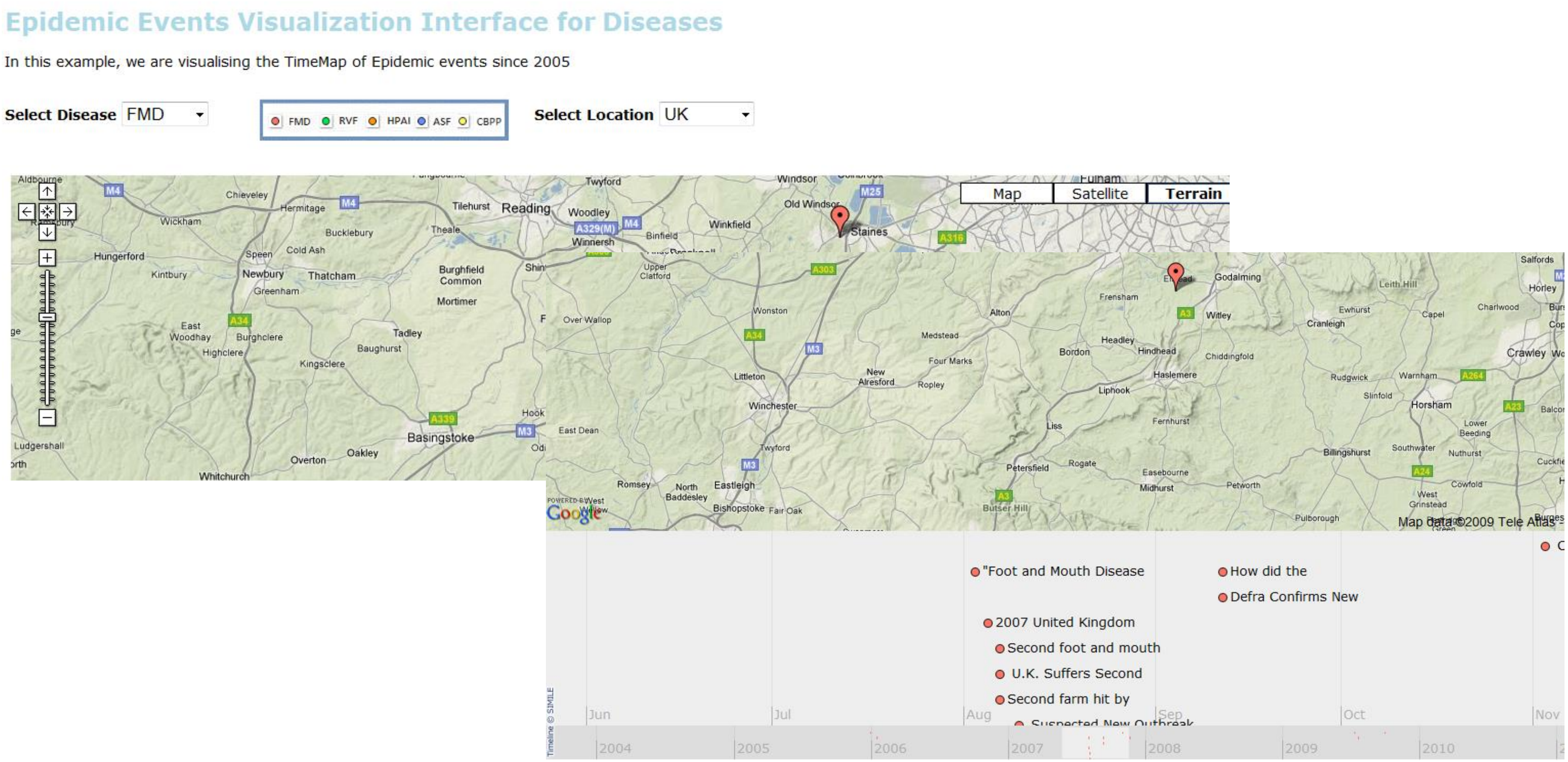
Link-Augmented Dynamic Topic Modeling

Saturday, 25 July 2015

William H. Hsu
Laboratory for Knowledge Discovery in Databases, Kansas State University
http://www.kddresearch.org

Acknowledgements
Northeastern University: Yizhou Sun
Stony Brook University: Leman Akoglu
Illinois: Jiawei Han
Kansas State: Wesam Elshamy. Surya Teja Kallumadi

Previous Work [1]: From Text to Spatiotemporal Events



http://fingolfin.user.cis.ksu.edu/timemap2gs

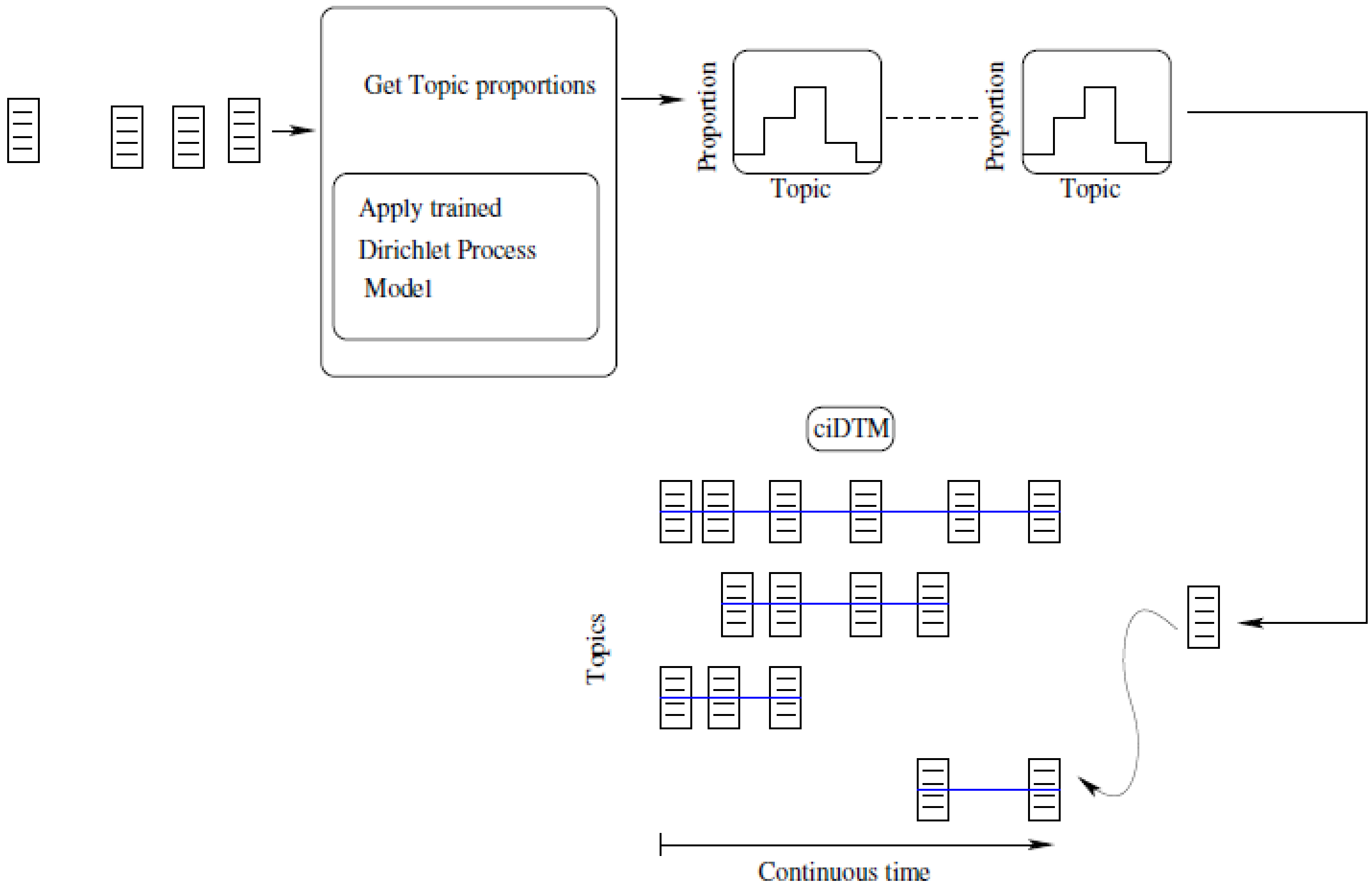
Based on

NLP Group NER Toolkit **© 2005-2010 Stanford University**

Simile **© 2003-2010 Massachusetts Institute of Technology**

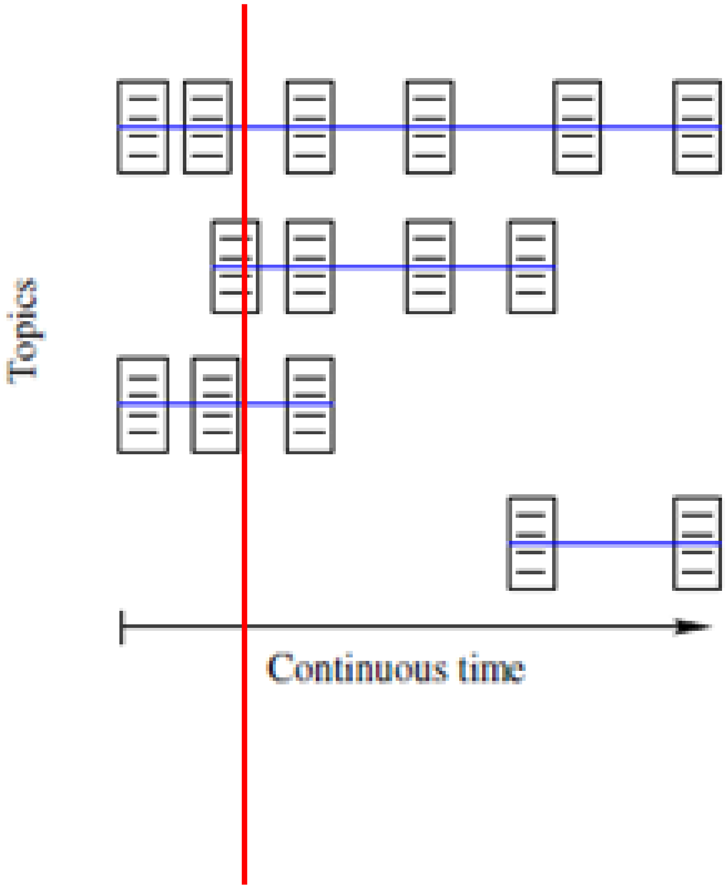
Google Maps **© 2007-2010 Tele Atlas, Inc. and Google, Inc.**

Previous Work [2]: Timeline Formation



Elshamy (2012)

Previous Work [3]: Simultaneous Topic Enumeration & Tracking (STET)



Time t: 3 extant topics

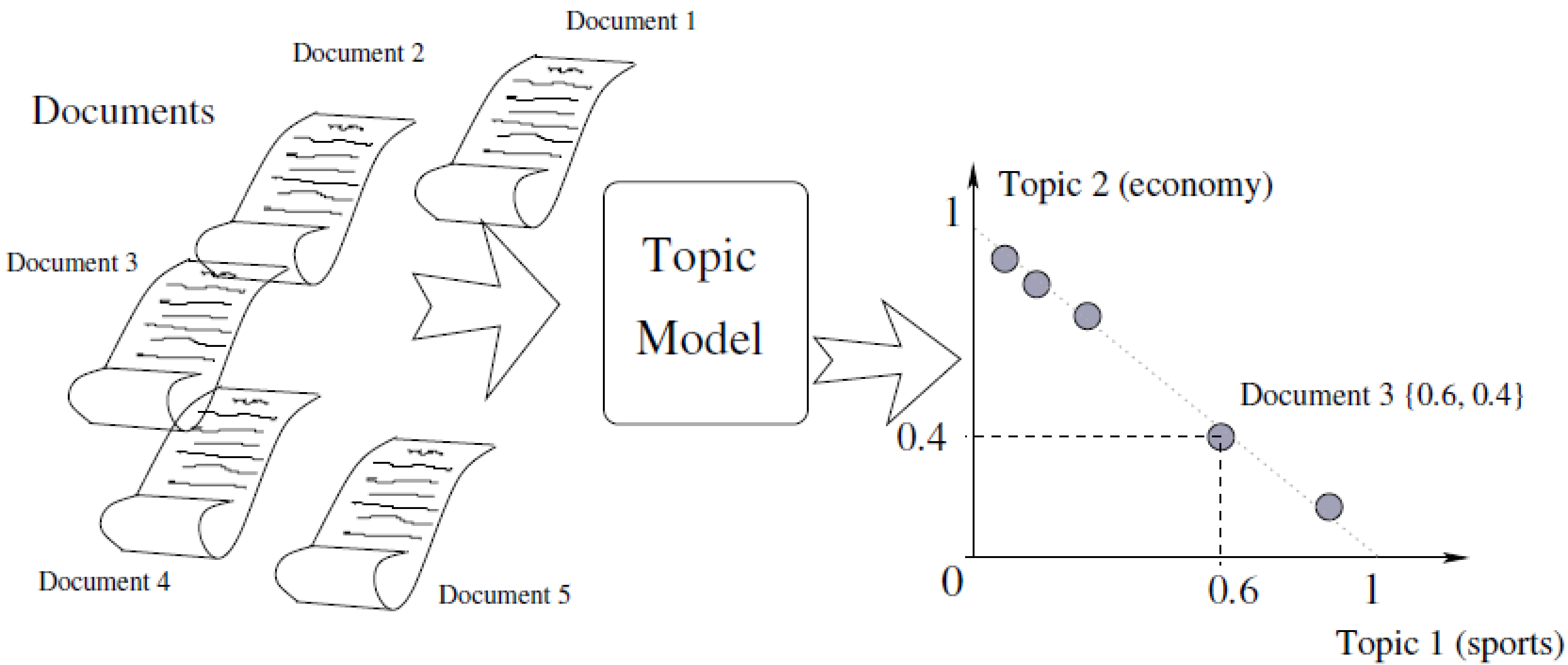
Time *t* + *k*: 2 extant topics

Adapted from Elshamy (2012)

Outline

- **Topic Modeling: Static (STM) vs. Dynamic (DTM)**
  - **Generative Model Concepts**
  - **Using Time in DTM**
- **Heterogeneous Information Network Analysis (HINA)**
- **From Link Analysis to Link-Augmented DTM**
  - **Community Detection: Gibson *et al.*, Lu & Getoor**
  - **From *RankClus* to *NetClus*: Sun *et al.***
  - ***LIMTopic* (Duan *et al.*) & Other Approaches**

Topic Modeling [1]: Basic Task (Static/Atemporal)



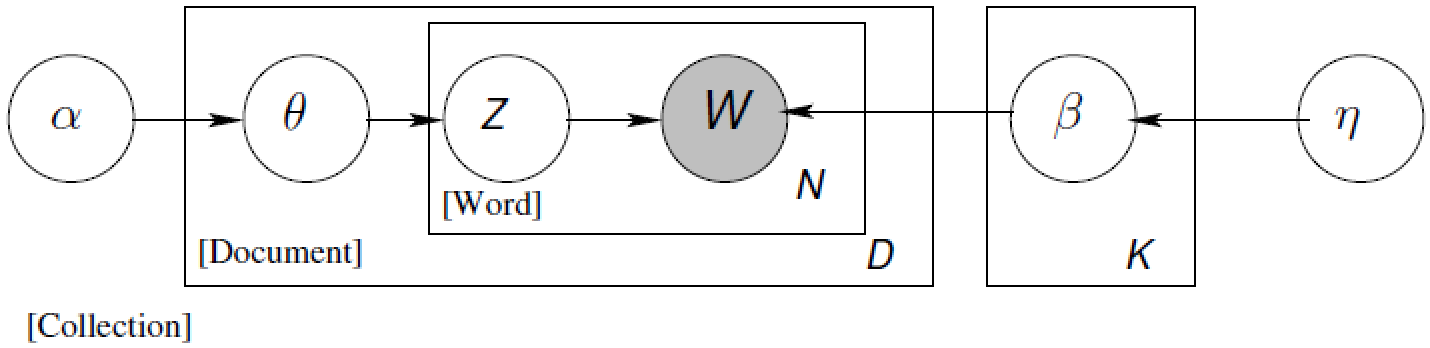
Probabilistic models to capture semantic structure of text

Elshamy (2012)

Outline

- <u>T</u>opic <u>M</u>odeling: <u>S</u>tatic (STM) vs. <u>D</u>ynamic (DTM)
  - Generative Model Concepts
  - Using Time in DTM
- <u>H</u>eterogeneous <u>I</u>nformation <u>N</u>etwork <u>A</u>nalysis (HINA)
- From Link Analysis to Link-Augmented DTM
  - Community Detection: Gibson *et al.*, Lu & Getoor
  - From *RankClus* to *NetClus*: Sun *et al.*
  - *LIMTopic* (Duan *et al.*) & Other Approaches

Topic Modeling [2]: Understanding Plate Notation



Jordan notation

N (5) : Number of words in a document.

D (1000) : Number of documents in a collection.

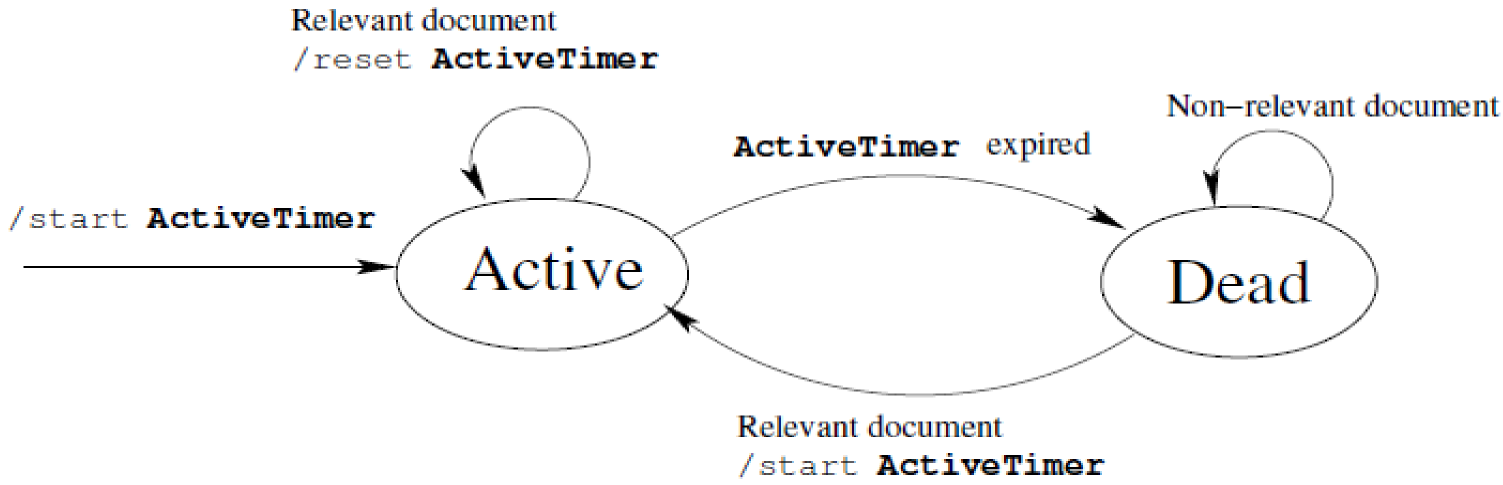
K (3) : Number of topics in the collection.

Adapted from Elshamy (2012)

Outline

- Topic Modeling: Static (STM) vs. Dynamic (DTM)
  - Generative Model Concepts
  - Using Time in DTM
- Heterogeneous Information Network Analysis (HINA)
- From Link Analysis to Link-Augmented DTM
  - Community Detection: Gibson *et al.*, Lu & Getoor
  - From *RankClus* to *NetClus*: Sun *et al.*
  - *LIMTopic* (Duan *et al.*) & Other Approaches

Dynamic Topic Modeling: HMM for Topic Detection & Tracking (TDT)



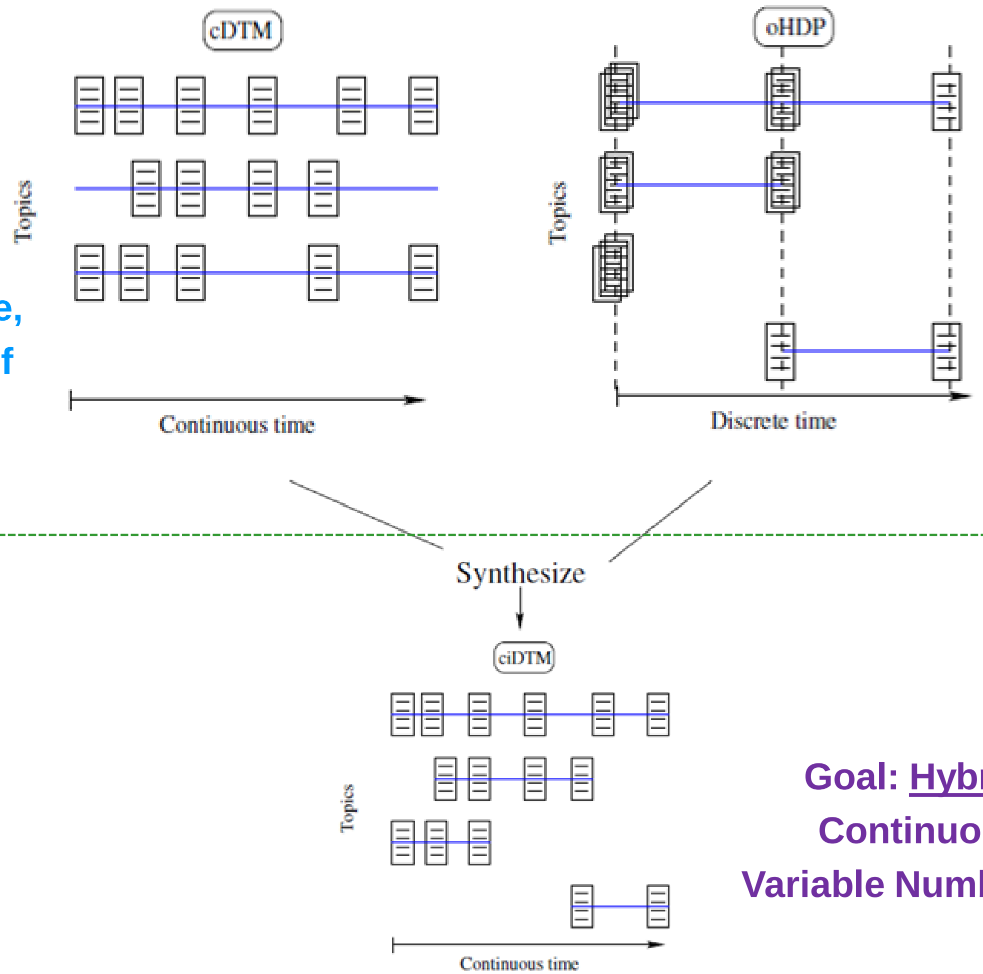
Hidden Markov Model (HMM)

- **Observable Transition: Topic Arrival (Detected by Topic Model)**
- **Hidden State: Topic Activity Level**
- **Problem: Topic Drift (Similar to Concept Drift)**

Adapted from Elshamy (2012)

Continuous Time *vs.* Variable Number of Topics

cDTM
Continuous Time,
Fixed Number of Topics



oHDP
Discrete Time,
Variable Number of Topics

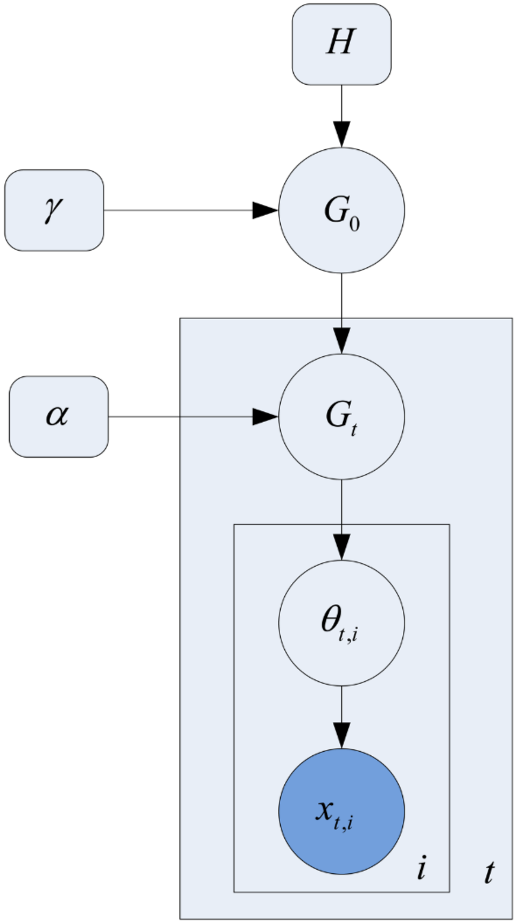
State of the Field

Goal: Hybrid Model
Continuous Time,
Variable Number of Topics

Elshamy (2012)

Hierarchical Dirichlet Process (HDP) Model

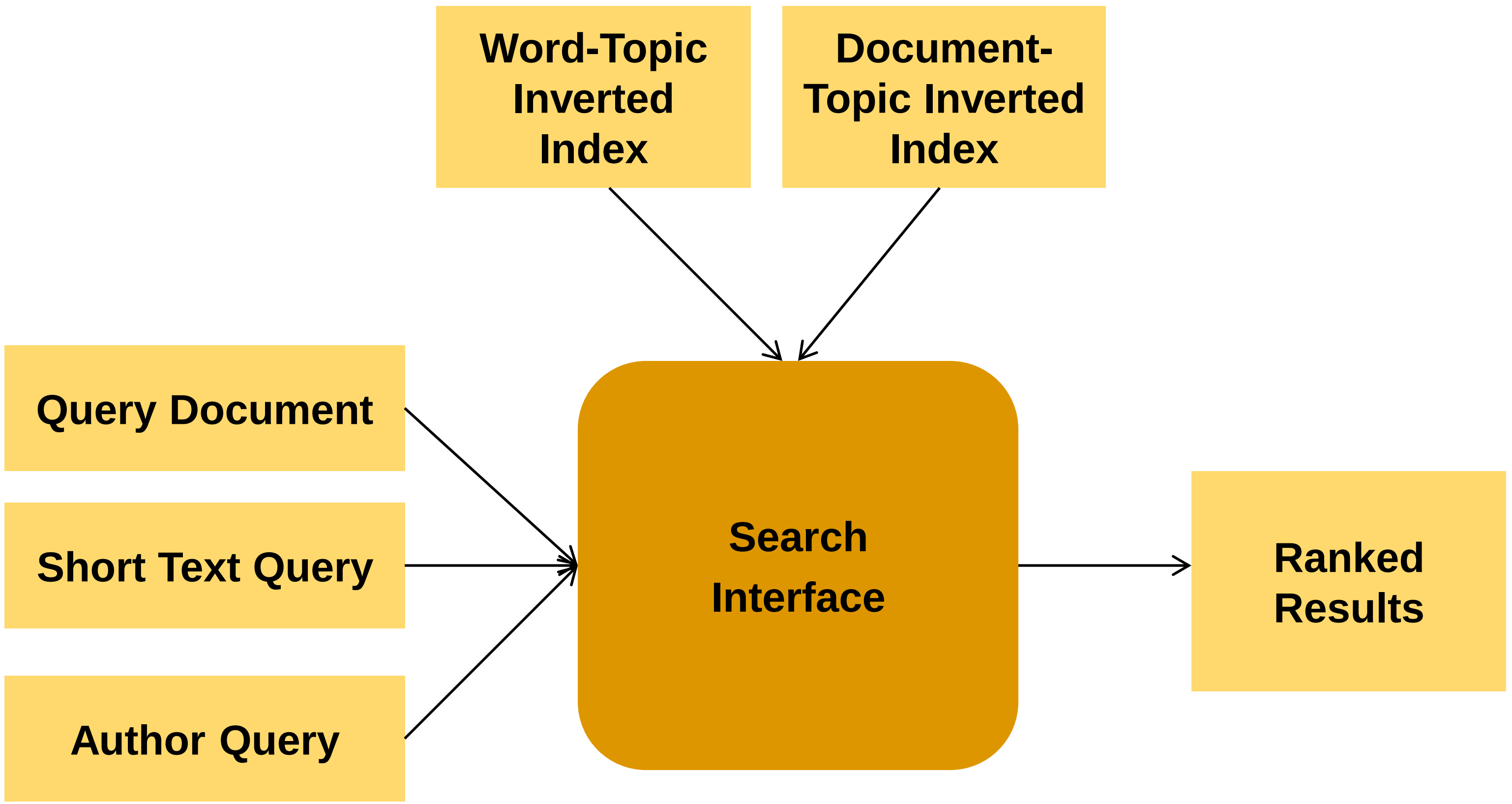
- Global prior G_0
- Prior G_t for words in document t
- Index i of words in document t
- Concentration parameter γ for G_0
- Concentration parameter α for G_t
- Topics $\theta_{t,i}$
- Words $x_{t,i}$



Fan *et al.* (2013)

Retrieving Documents

Corpus representation from topic model



Information Retrieval Task [1]: Overview

- Representation
  - Document *d*: bag of words
  - Author *a*: mixture of all (co-)authored documents
  - Query *q*: (usually smaller) bag of words
- Process: given input query *q*
  - Intake: perform data processing on incoming *q*
  - Infer topic mixture for *q* using trained topic model
  - Use to retrieve documents, authors for this *q*

Information Retrieval Task [2]: Static Setting

- Given: document d, query q
- Estimate using query likelihood model, $P(q \mid d)$
- Maintain two inverted indices
  - $P(w \mid k)$ – word w given topic k
  - $P(k \mid d)$ – topic k given document d
- Use these indices for fast retrieval
- Tested on small corpus: 100 documents

Information Retrieval Task [3]: Dynamic Setting

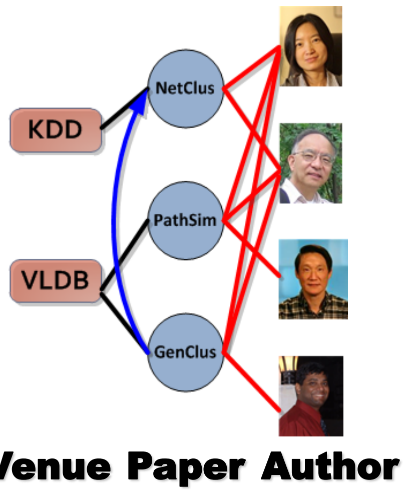
- Similar to static setting
- Inputs: query, time duration (search window)
- Query likelihood model: $P(q \mid d)$
- For each time stamp in search window:
  - Maintain two inverted indices as in static setting
  - Perform query-based retrieval at each time stamp
  - Aggregate results
- Currently under development

OUTLINE

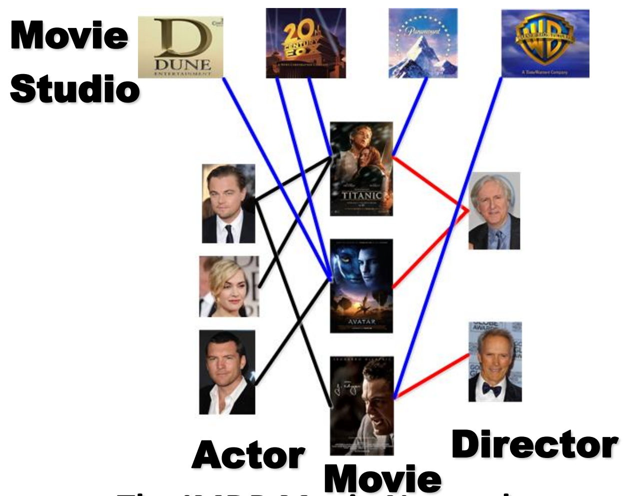
- **Topic Modeling: Static (STM) vs. Dynamic (DTM)**
  - **Generative Model Concepts**
  - **Using Time in DTM**
- **Heterogeneous Information Network Analysis (HINA)**
- **From Link Analysis to Link-Augmented DTM**
  - **Community Detection: Gibson *et al.*, Lu & Getoor**
  - **From *RankClus* to *NetClus*: Sun *et al.***
  - ***LIMTopic* (Duan *et al.*) & Other Approaches**

Heterogeneous Information Network Analysis (HINA)

Real-world data: Multiple object types and/or multiple link types



DBLP Bibliographic Network



The IMDB Movie Network

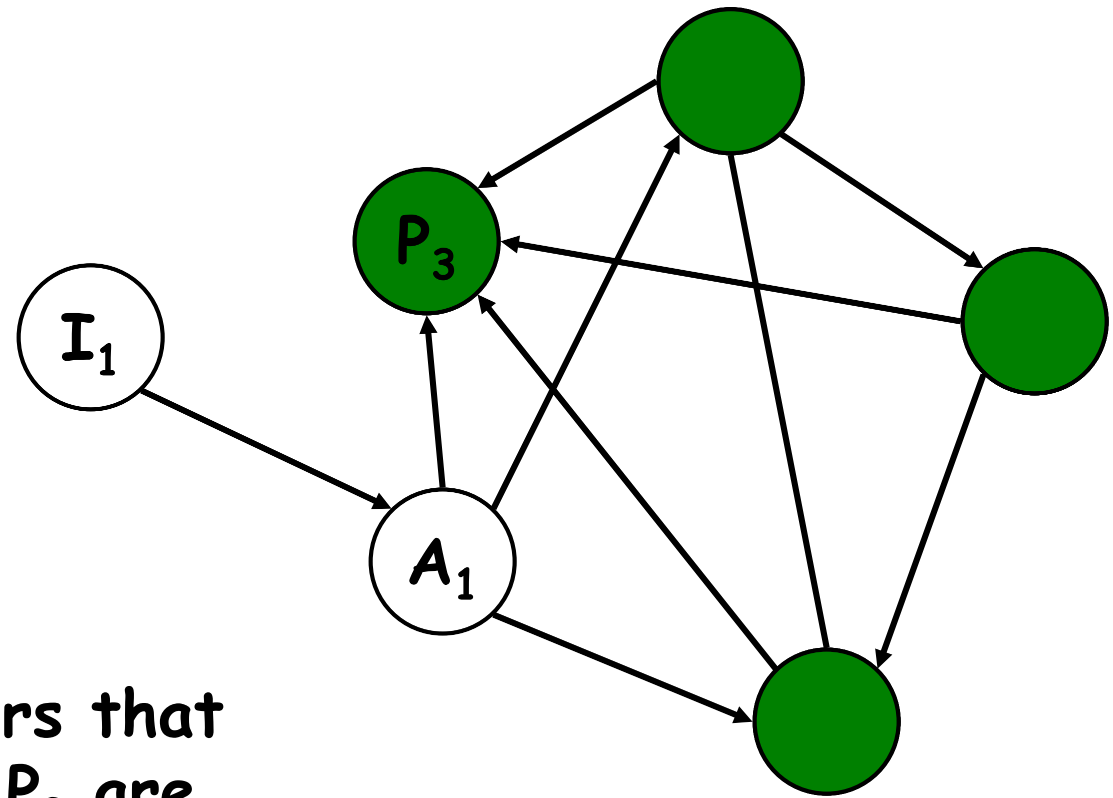

The Facebook Network

Han, Wang, & El-Kishky (KDD 2014)

Outline

- Topic Modeling: Static (STM) vs. Dynamic (DTM)
  - Generative Model Concepts
  - Using Time in DTM
- Heterogeneous Information Network Analysis (HINA)
- From Link Analysis to Link-Augmented DTM
  - Community Detection: Gibson *et al.*, Lu & Getoor
  - From *RankClus* to *NetClus*: Sun *et al.*
  - *LIMTopic* (Duan *et al.*) & Other Approaches

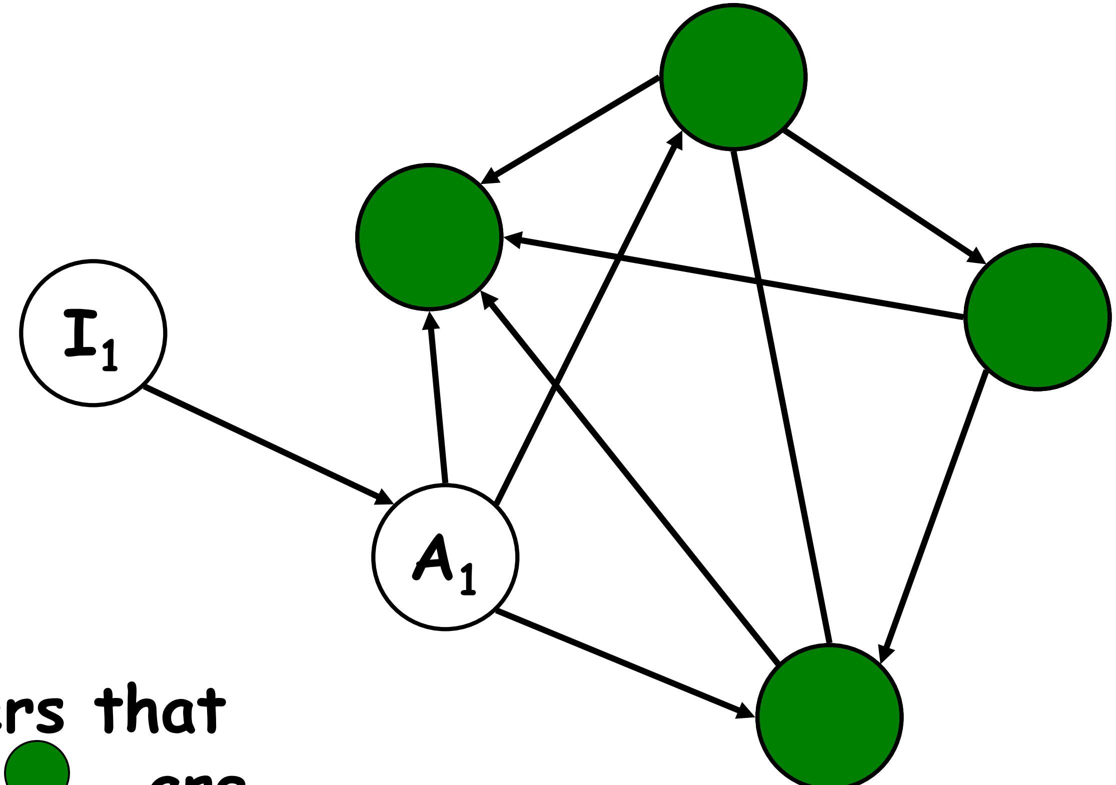
Link Analysis [1]: Instance-based Dependencies



Papers that cite P_3 are likely to be

Getoor, Bhattacharya, Lu, & Sen (MRDM 2004)

Link Analysis [2]: Class-based Dependencies



Papers that cite ● are likely to be ●

Getoor, Bhattacharya, Lu, & Sen (MRDM 2004)

Outline

- Topic Modeling: Static (STM) vs. Dynamic (DTM)
  - Generative Model Concepts
  - Using Time in DTM
- Heterogeneous Information Network Analysis (HINA)
- From Link Analysis to Link-Augmented DTM
  - Community Detection: Gibson *et al.*, Lu & Getoor
  - From *RankClus* to *NetClus*: Sun *et al.*
  - *LIMTopic* (Duan *et al.*) & Other Approaches

Link Descriptions

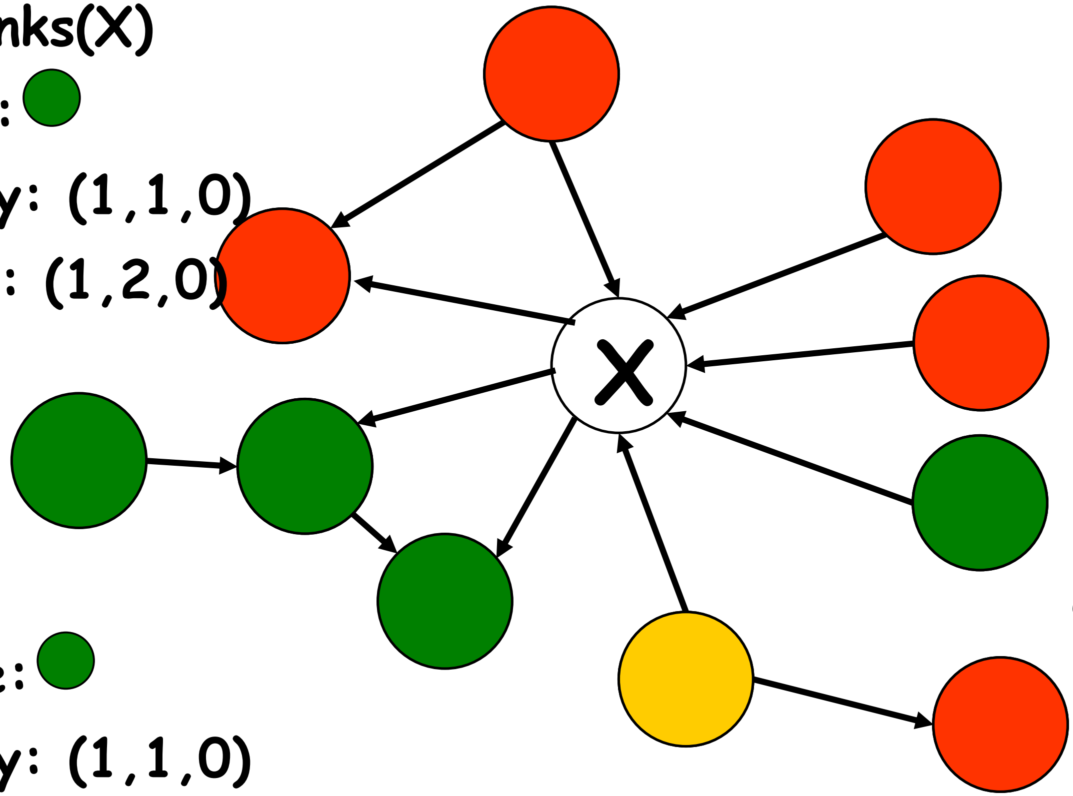
Categories

Out-Links(X)

mode:

binary: (1,1,0)

count: (1,2,0)



In-Links(X)

mode:

binary: (1,1,1)

count: (3,1,1)

CO(X)

mode:

binary: (1,1,0)

count: (2,1,0)

CI(X)

mode:

binary: (1,0,0)

count: (2,0,0)

Getoor, Bhattacharya, Lu, & Sen (MRDM 2004)

Predictive Model for Classification

- A structured logistic regression
  - Compute P(c | OA(X)) and P(c | LD(X)) separately using separate logistic regression models

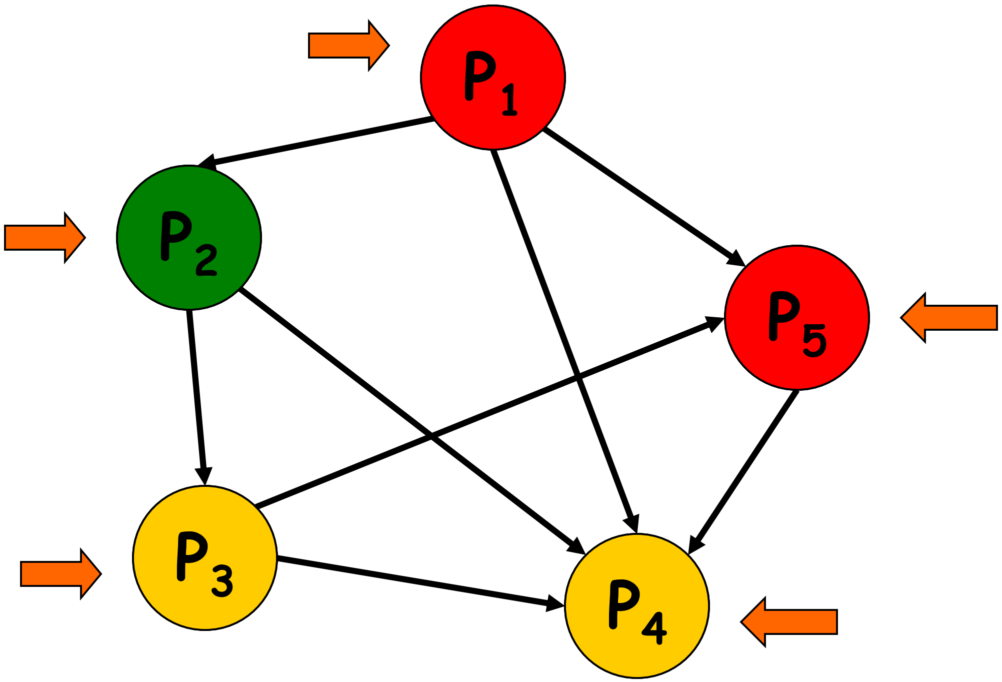
$$\hat{c}(x) = \arg\max_{c} P(c \mid OA(x)) \prod_{f \in \{In, Out, CI, CO\}} \frac{P(c \mid LD_f(x))}{P(c)}$$

  - where OA(X) are the object attributes and $LD_f(X)$ are the link features

Getoor, Bhattacharya, Lu, & Sen (MRDM 2004)

PREDICTION

- category set { ● ● ● }



Step 1: Bootstrap using object attributes only

Getoor, Bhattacharya, Lu, & Sen (MRDM 2004)

Outline

- Topic Modeling: Static (STM) vs. Dynamic (DTM)
  - Generative Model Concepts
  - Using Time in DTM
- Heterogeneous Information Network Analysis (HINA)
- From Link Analysis to Link-Augmented DTM
  - Community Detection: Gibson *et al.*, Lu & Getoor
  - From *RankClus* to *NetClus*: Sun *et al.*
  - *LIMTopic* (Duan *et al.*) & Other Approaches

NetClus [1]

- Algorithm to perform clustering over heterogeneous network
- Performs iterative ranking of clustering of objects.
- Probabilistic generative model is used to model probability of generating different objects from each cluster
- Maximum likelihood technique used to evaluate posterior probability of presence of object in cluster

Gupta, Aggarwal, Han, Sun (ASONAM 2011)

NetClus [2]

- Priors: Initialize prior probabilities $\{P(o|ck)\}_{k=1}^{K}$
- Initialize: Generate initial net-clusters. $\{c_k^0\}_{k=1}^{K}$
- Rank: Build probabilistic generative model for each net-cluster, i.e., $\{P(o|c_k^t)\}_{k=1}^{K}$
- Cluster-target: Compute p$(c_k^t|o)$ for target objects and adjust their cluster assignments.
- Iterate: Repeat steps 3 and 4 until the clusters don't change significantly.
- Cluster-attribute: Calculate p$(c_k^*|o)$ for each attribute object in each net-cluster.
- Return p$(c_k^*|o)$

Gupta, Aggarwal, Han, Sun (ASONAM 2011)

Using NetClus

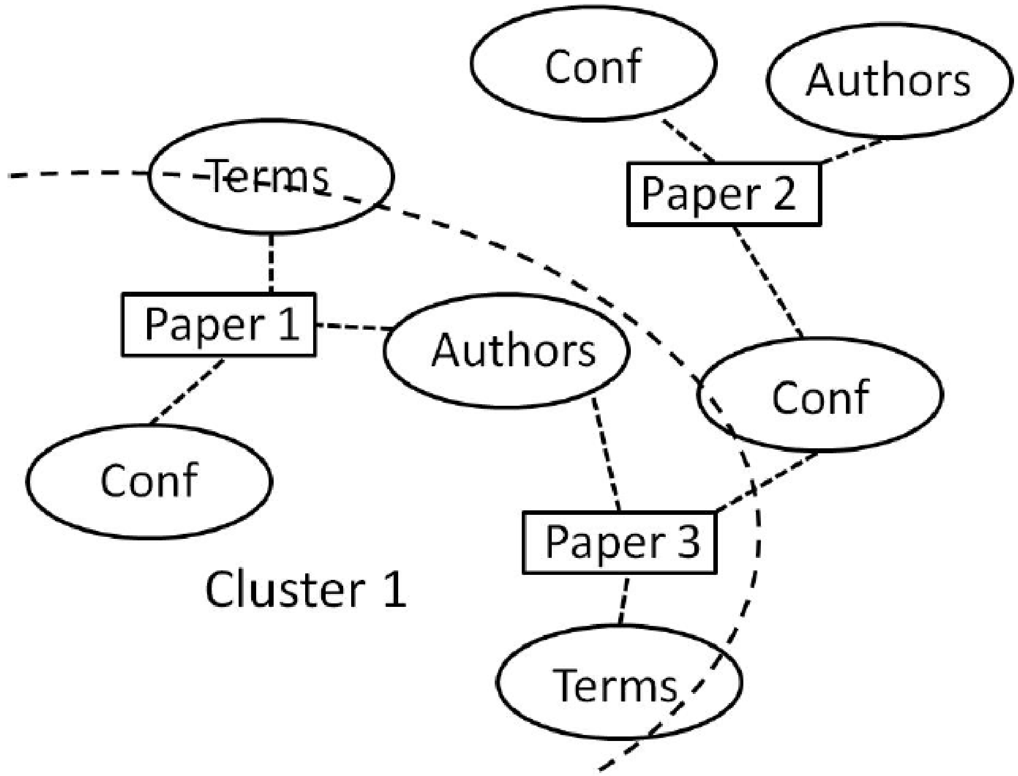
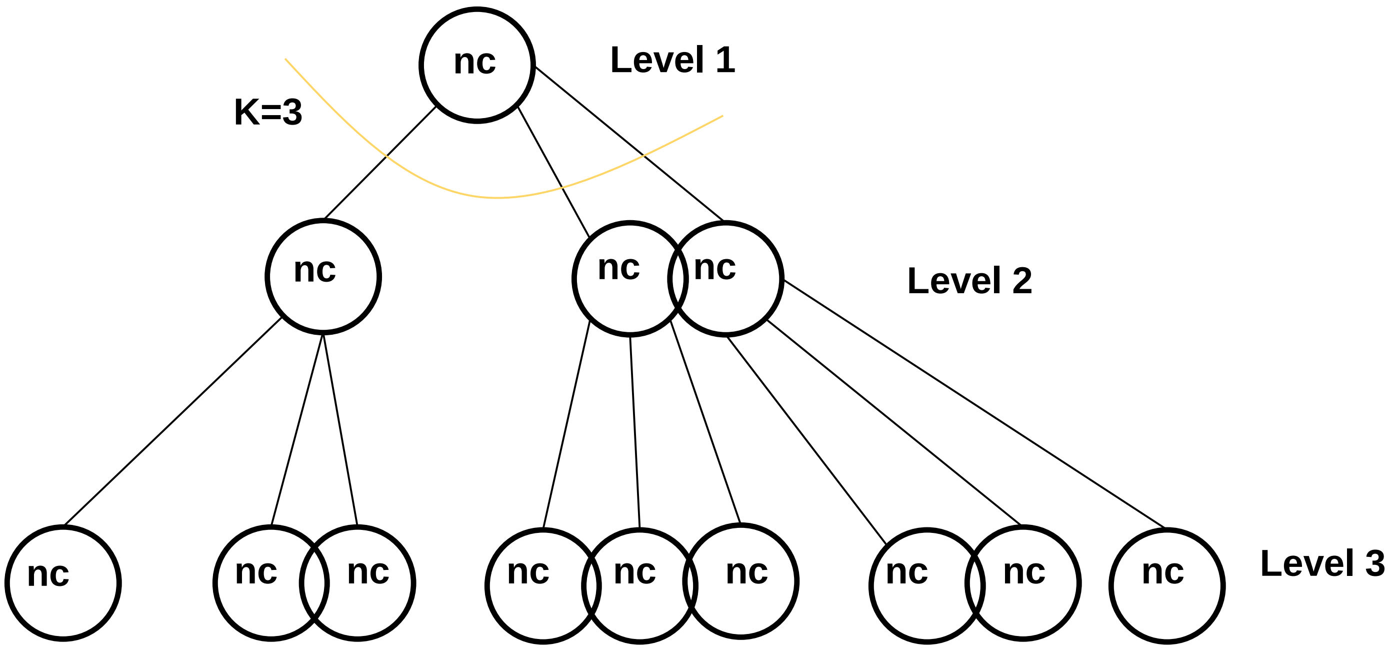


Figure 1: Net-cluster view of DBLP network

Gupta, Aggarwal, Han, Sun (ASONAM 2011)

NetClus Results: Net Cluster Tree



Gupta, Aggarwal, Han, Sun (ASONAM 2011)

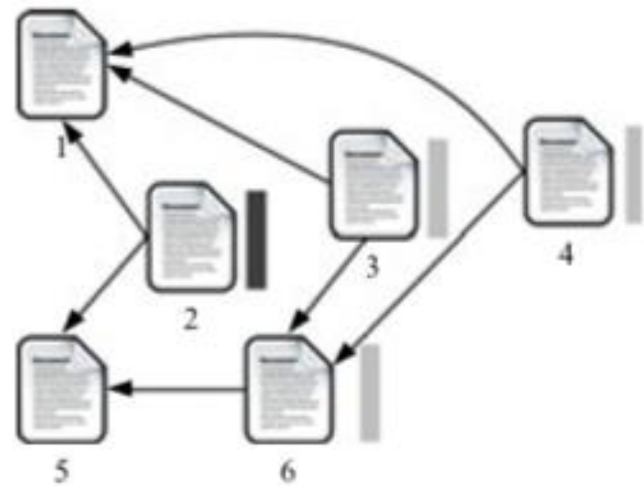
Outline

- **Topic Modeling: Static (STM) vs. Dynamic (DTM)**
  - **Generative Model Concepts**
  - **Using Time in DTM**
- **Heterogeneous Information Network Analysis (HINA)**
- **From Link Analysis to Link-Augmented DTM**
  - **Community Detection: Gibson *et al.*, Lu & Getoor**
  - **From *RankClus* to *NetClus*: Sun *et al.***
  - ***LIMTopic* (Duan *et al.*) & Other Approaches**

LIMTopic [1]: Document Networks

word \ doc	1	2	3	4	5	6
normalized	0	0	1	1	0	1
cuts	0	0	1	1	0	1
image	0	0	1	1	0	1
segmentation	0	0	1	1	0	1
graph	0	0	1	1	0	1
spectral	0	1	0	0	0	0
clustering	0	1	0	0	0	0
community	0	1	0	0	0	0
detection	0	1	0	0	0	0
network	0	1	0	0	0	0

(a) Documents



(b) Link structure

Fig. 1. An artificial document network. There are two topics in these documents, which are represented by gray and dark bars respectively. Documents are labeled by corresponding bars beside them.

Duan, Li, Li, Zhang, Gu, & Wen (IEEE TKDE 2011)

LIMTopic [2]: PageRank-Like Ranking Algorithms

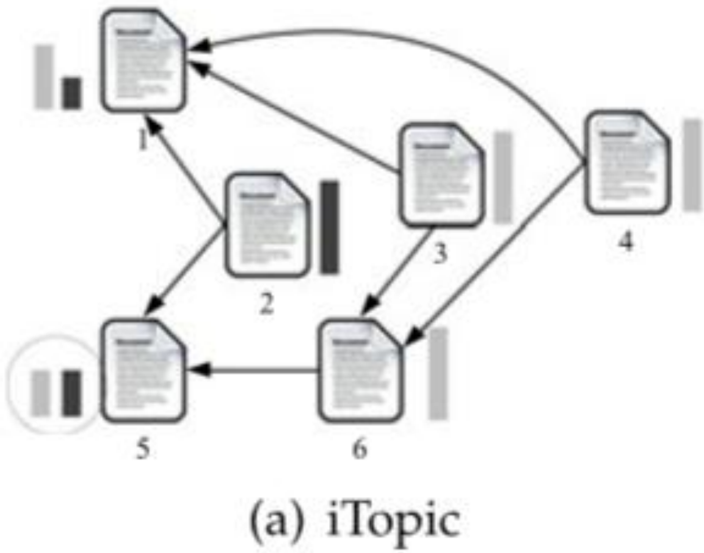


Fig. 2. The topic proportions of documents output by iTopic and RankTopic respectively. Higher bars indicate more proportions.

Duan, Li, Li, Zhang, Gu, & Wen (IEEE TKDE 2011)

LIMTopic [3]: Link-Augmented STM

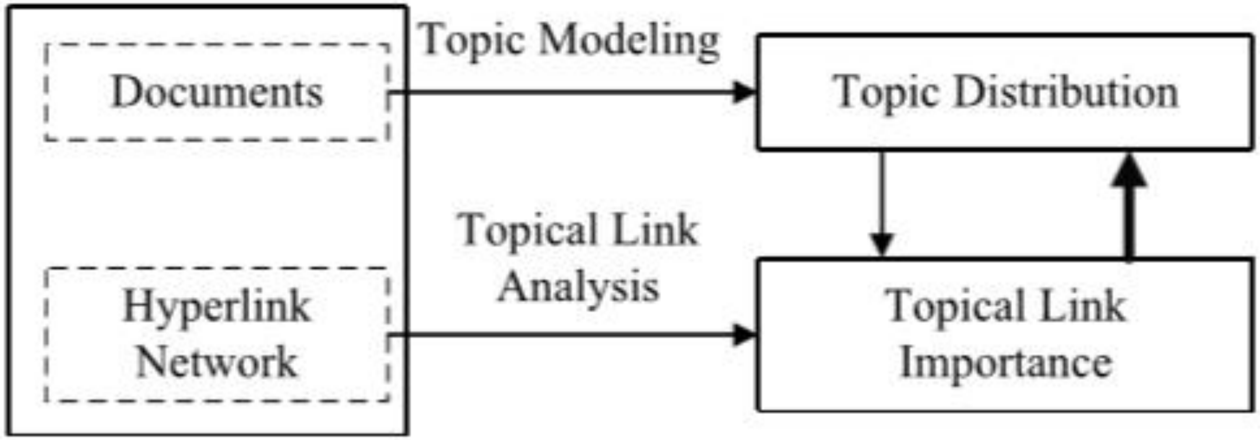


Fig. 3. Mutual enhancement framework between topic distribution and link based importance.

Duan, Li, Li, Zhang, Gu, & Wen (IEEE TKDE 2011)

LIMTopic [4]: Using Link Importance

Link importance

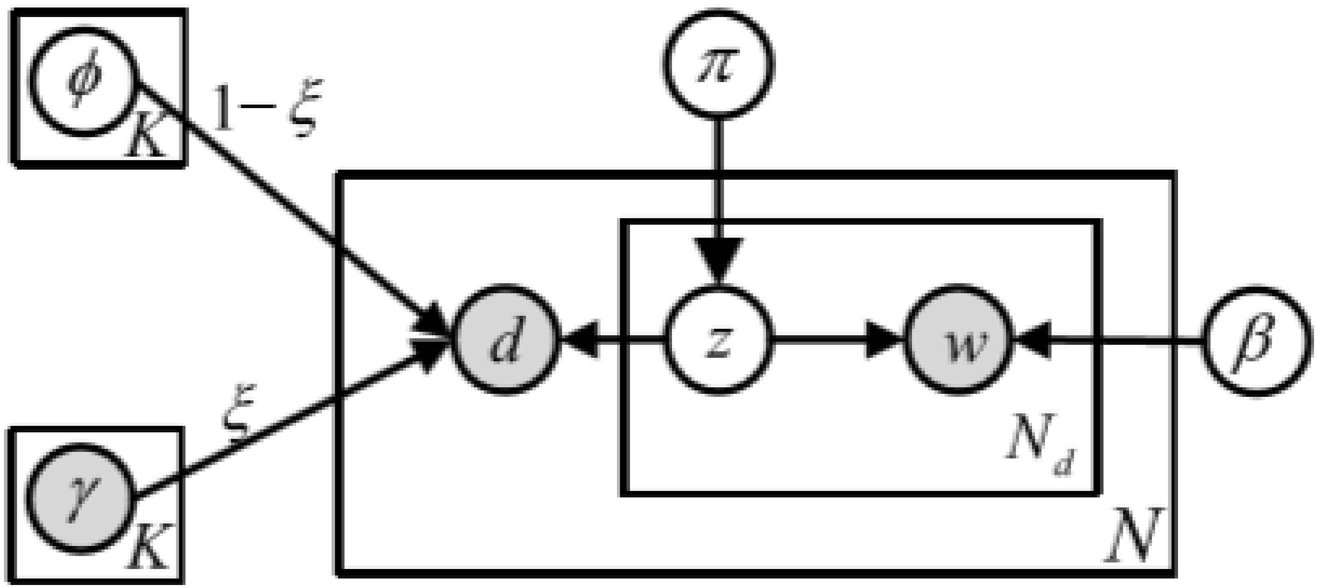
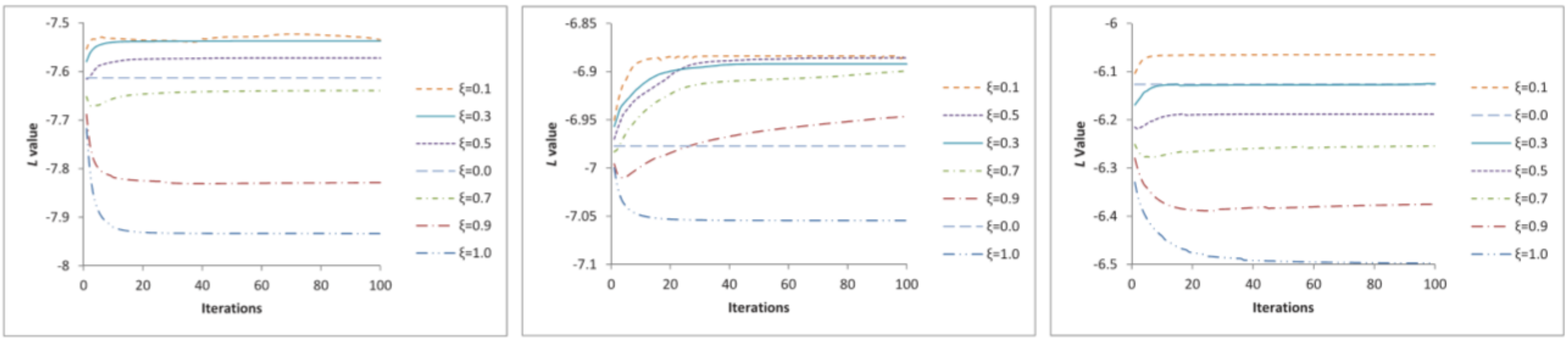


Fig. 4. Link importance based topic modeling framework.

Duan, Li, Li, Zhang, Gu, & Wen (IEEE TKDE 2011)

LIMTopic [5]: Evaluating Topic Models for Ranking



(a) RankTopic on ArnetMiner (b) RankTopic on Citeseer (c) HITSTopic on Cora

Fig. 5. L curves of LIMTopic with the iterations.

Duan, Li, Li, Zhang, Gu, & Wen (IEEE TKDE 2011)