Topic Modeling with Network Regularization

Qiaozhu Mei, Deng Cai, Duo Zhang, ChengXiang Zhai

University of Illinois at Urbana-Champaign

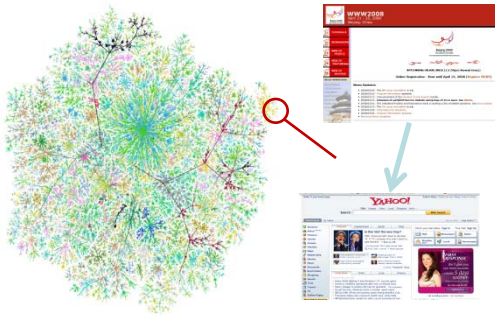
Outline

- Motivation
- An optimization framework
  - probabilistic topic model with graph regularization
- NetPLSA
- Experiments
- Summary

Text Collections with Network Structure

Blog articles + friend network

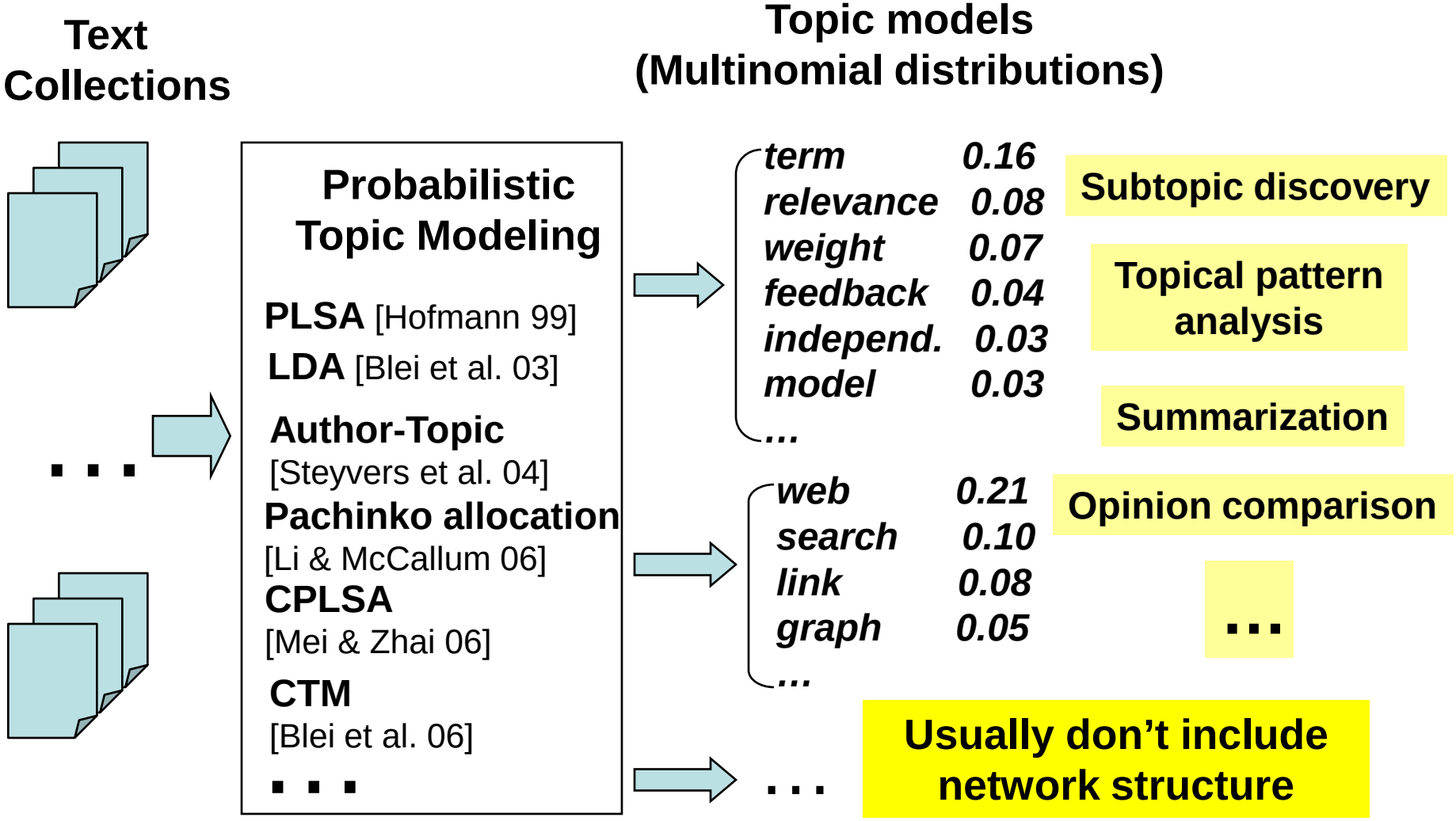
News + geographic network



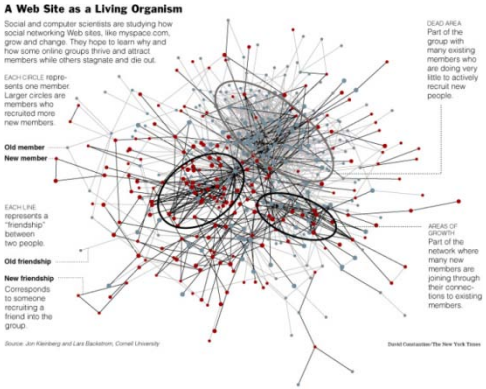
Web page + hyperlink structure

- Literature + coauthor/citation network
- Email + sender/receiver network
- …

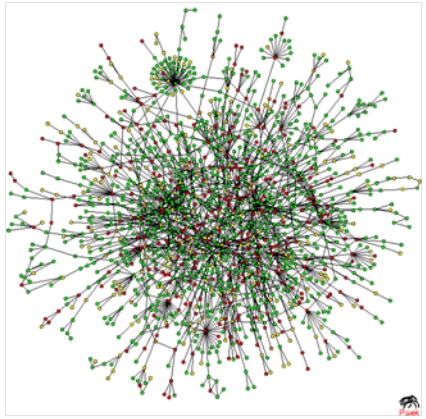
Probabilistic Topic Models for Text Mining



Social Network Analysis



- Kleinberg and Backstrom 2006, New York Times



- Jeong et al. 2001 Nature 411

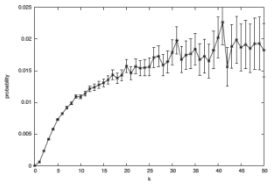
Generation, evolution
e.g., [Leskovec 05]

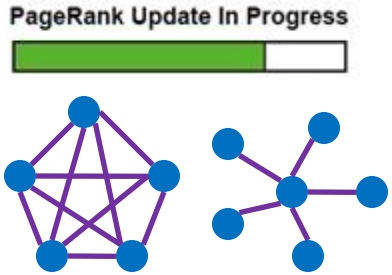
Community extraction
e.g., [Kleinberg 00];

Diffusion
[Gruhl 04]; [Backstrom 06]

Ranking
e.g., [Brin and Page 98]; [Kleinberg 98]

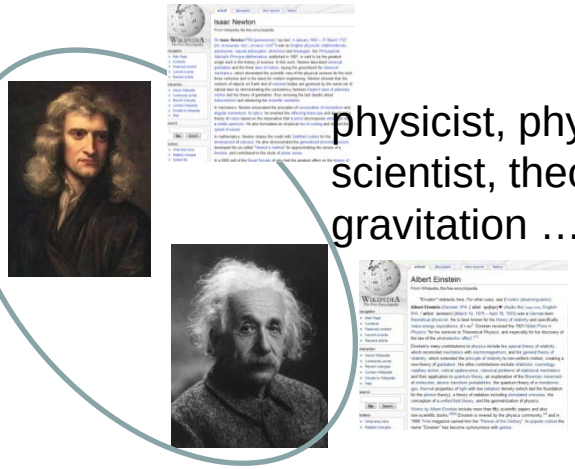
Structural patterns
e.g., [Yan 02]





Usually don't model topics in text

Importance of Topic Modeling Plus Network Analysis

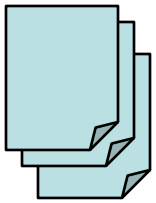


physicist, physics, scientist, theory, gravitation …

writer, novel, best-sell, book, language, film…

Topic modeling to help community extraction

Network analysis to help topic extraction



Computer Science Literature

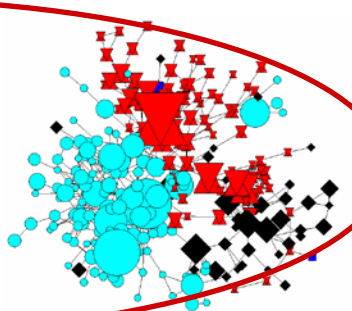
=

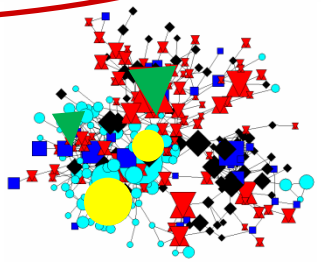
Information Retrieval +
Data Mining +
Machine Learning, …

or

Domain Review +
Algorithm +
Evaluation, …

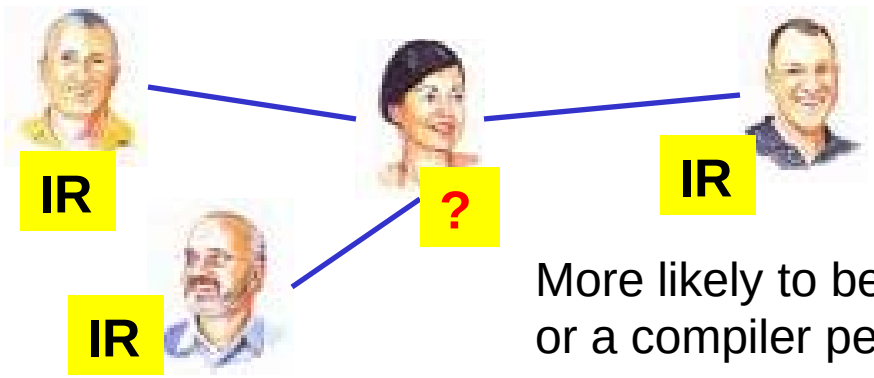
?





Intuitions

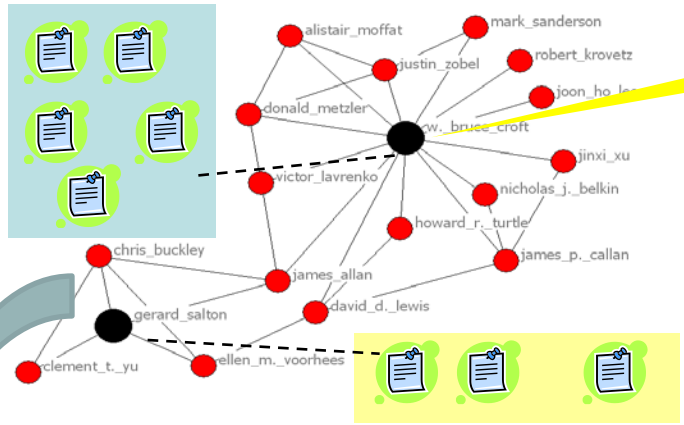
- People working on the same topic belong to the same "topical community"
- Good community: coherent topic + well connected
- A topic is semantically coherent if people working on this topic also collaborate a lot



Intuition: my topics are similar to my neighbors

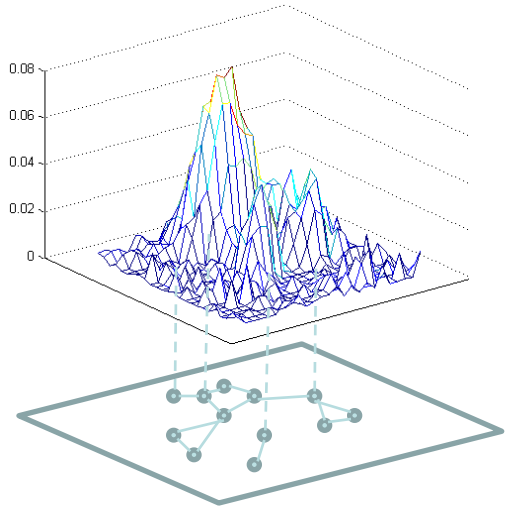
More likely to be an IR person or a compiler person?

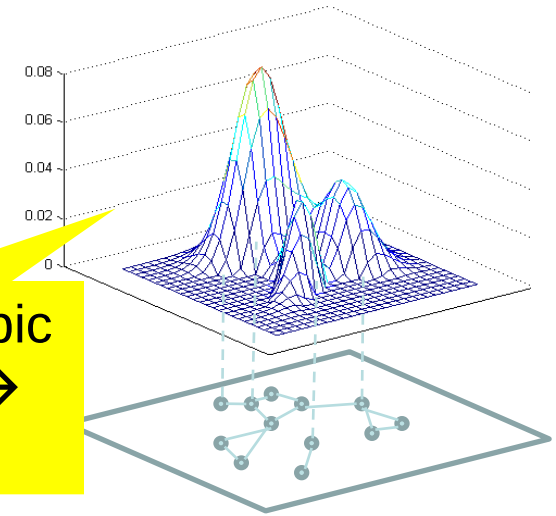
Social Network Context for Topic Modeling



e.g. coauthor network

- Context = author
- Coauthor = similar contexts
- Intuition: I work on similar topics to my neighbors





Smoothed Topic distributions → $P(\theta_j|\text{author})$

Challenging Research Questions

- How to formalize the intuitive assumption?
  - smoothed topic distributions over neighbors
  - without hurting topic modeling
- How to map a topic on a network structure?
- How to interpret the semantics of communities on a network?
  - These vertices form a community, but why?
  - Topical Communities

A Unified Optimization Framework

- Probabilistic topic modeling as an optimization problem (e.g., PLSA/LDA: Maximum Likelihood):

$$O(Collection \mid Model) = \log(P(Collection \mid Model))$$

- Regularized objective function with network constrains
  - Topic distribution are smoothed over adjacent vertices

$$O(Collection, Network \mid Model)$$
$$= \log(P(Collection \mid Model)) \oplus Regularizer(Model, Network)$$

$$ModelParams = \arg\max_{params} O(Collection[, Network] \mid Model)$$

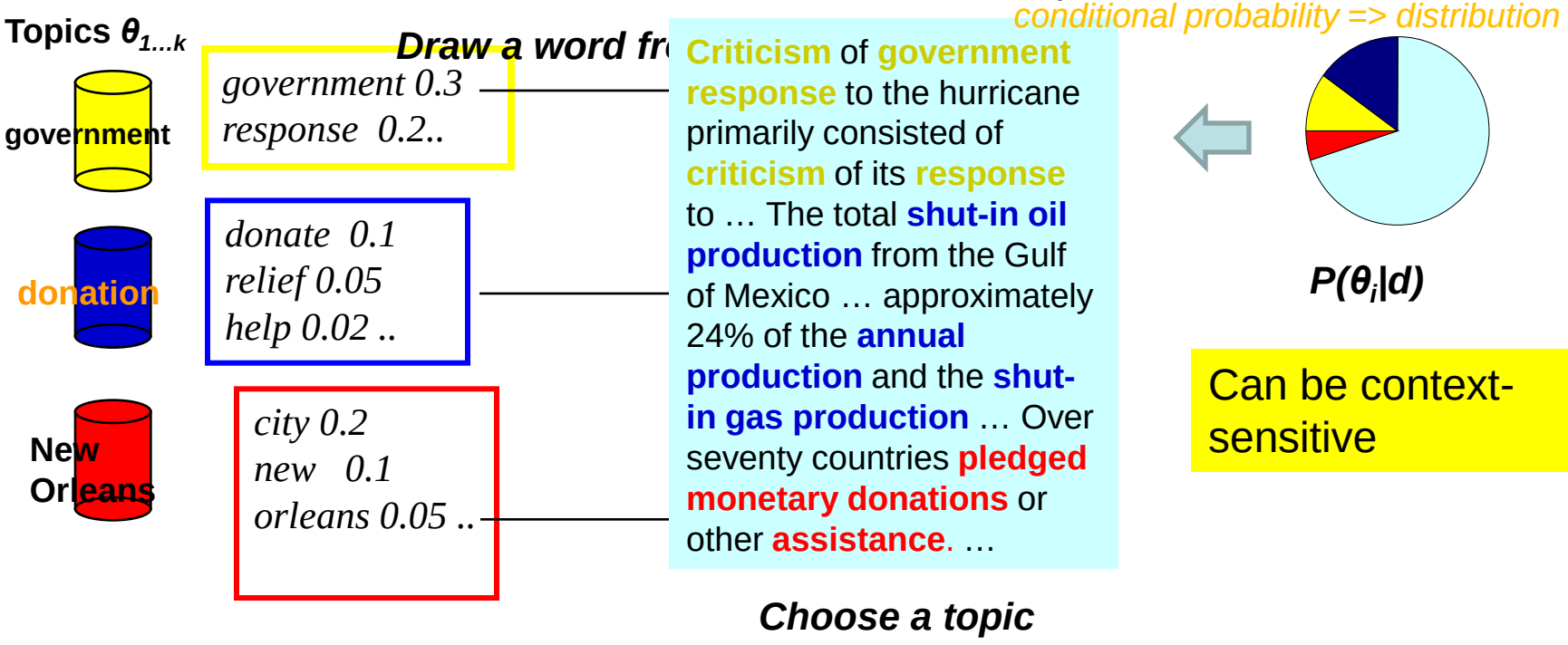
PLSA : Probabilistic Latent Semantic Analysis (Hofmann '99)

- Generation Process:

c:the occurrences of word in doc

$$O(C) = L(C) = \sum_d \sum_w c(w,d) \log \sum_{j=1}^{k} p(\theta_j \mid d) p(w \mid \theta_j))$$

conditional probability => distribution



Can be context-sensitive

NetPLSA

- Basic Assumption: Neighbors have similar topic distribution

PLSA

topic distribution of a document

$$O(C,G) = (1-\lambda)\cdot(\sum_d \sum_w c(w,d)\log \sum_{j=1}^{k} p(\theta_j \mid d)p(w \mid \theta_j))$$

$$+\lambda\cdot(-\frac{1}{2}\sum_{\langle u,v\rangle\in E} w(u,v)\sum_{j=1}^{k}(p(\theta_j \mid u)-p(\theta_j \mid v))^2)$$

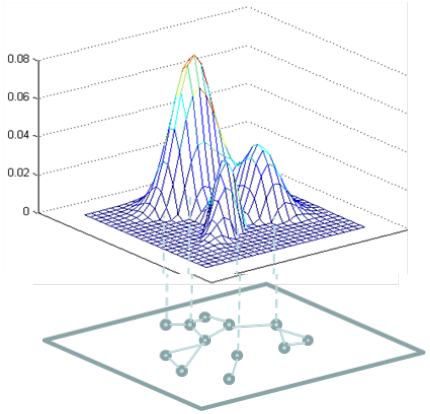
tradeoff

difference of topic distribution

importance (weight) of an edge

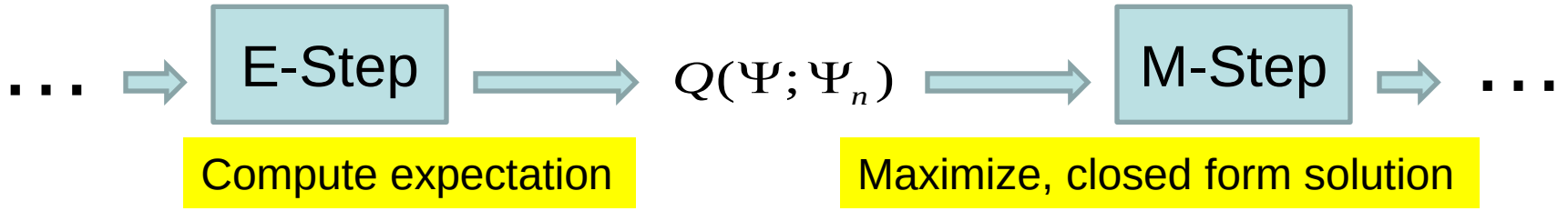
Graph Harmonic Regularizer, Generalization of [Zhu '03],

$$=\frac{1}{2}\sum_{j=1...k} f_j^T \Delta f_j \quad ,where \quad f_{j,u} = p(\theta_j \mid u)$$

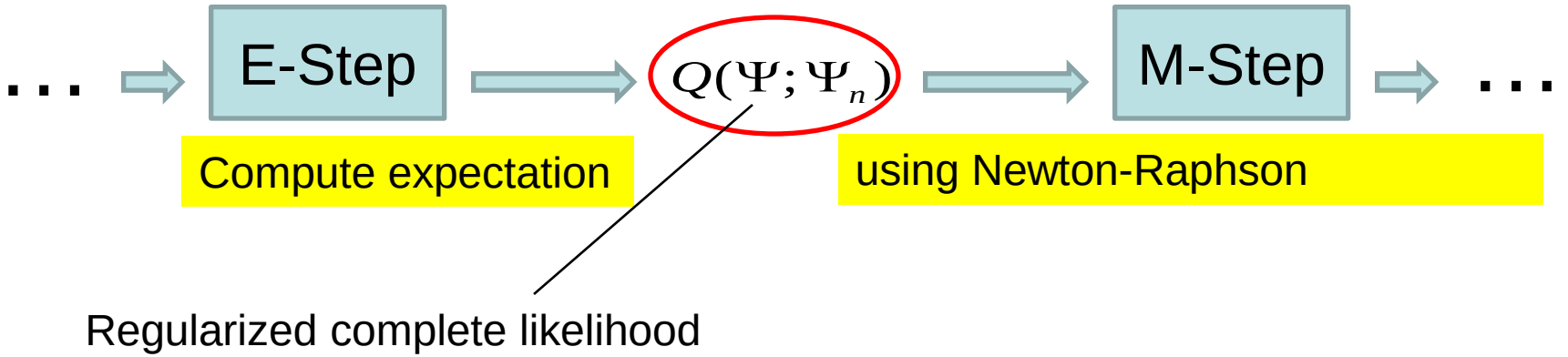


Parameter Estimation

- PLSA: EM algorithm

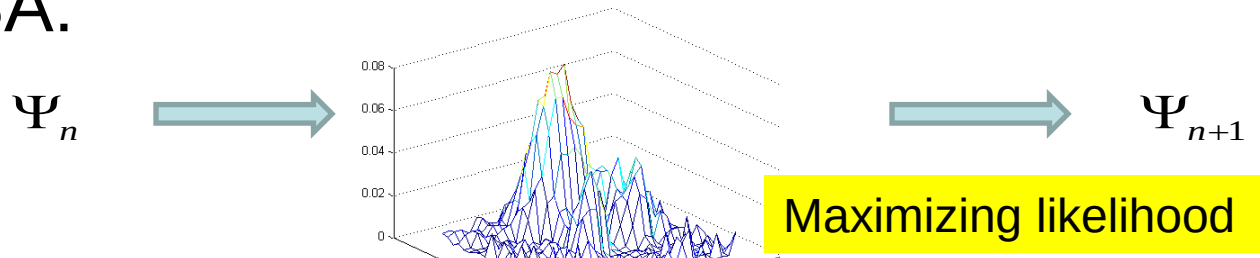


- NetPLSA: Generalized EM Algorithm

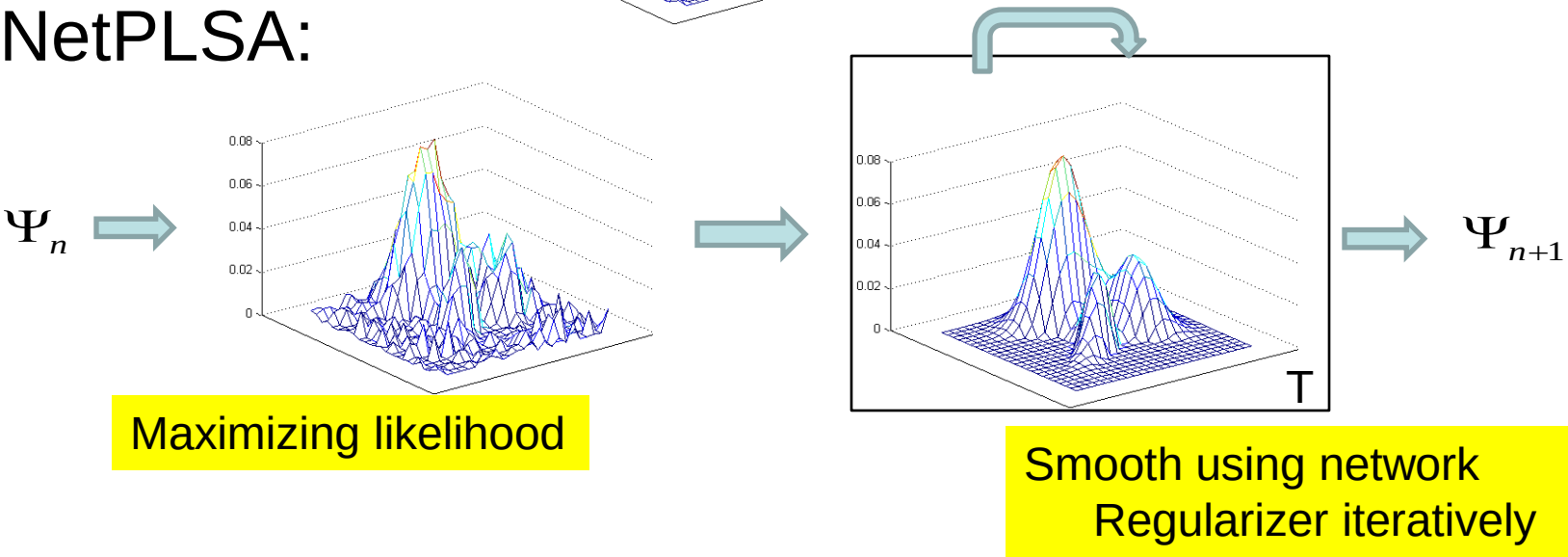


How it Works in M Step

- PLSA:

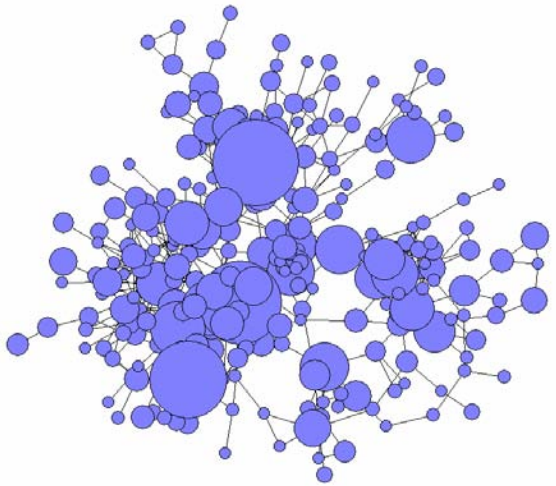


- NetPLSA:



Experiments

- Bibliography data and coauthor networks
  - DBLP: text = titles; network = coauthors
  - Four conferences (expect 4 topics): SIGIR, KDD, NIPS, WWW

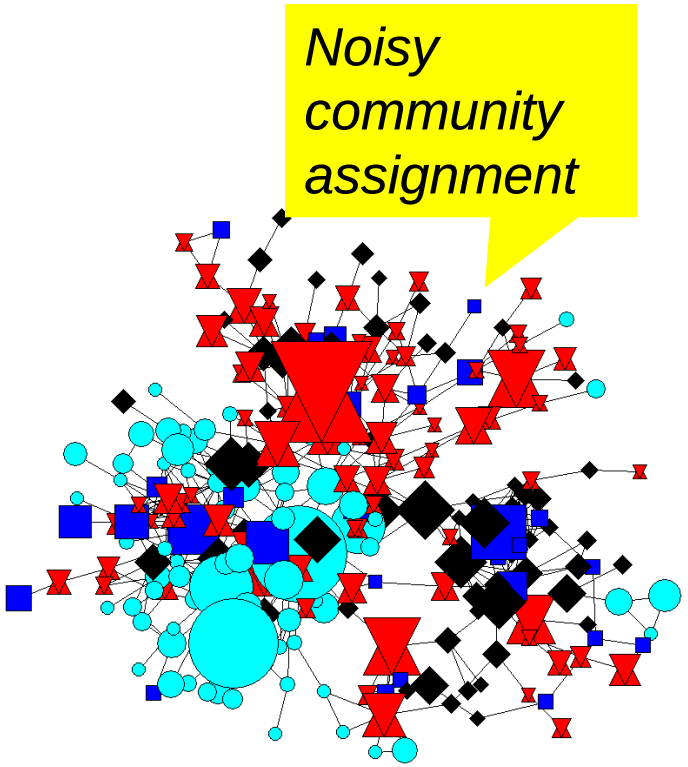


- Blog articles and Geographic network
  - Blogs from spaces.live.com containing topical words, e.g. “weather”
  - Network: US states (adjacent states)

Topical Communities with PLSA

Topic 1		Topic 2		Topic 3		Topic 4	
term	0.02	peer	0.02	visual	0.02	interface	0.02
question	0.02	patterns	0.01	analog	0.02	towards	0.02
protein	0.01	mining	0.01	neurons	0.02	browsing	0.02
training	0.01	clusters	0.01	vlsi	0.01	xml	0.01
weighting	0.01	stream	0.01	motion	0.01	generation	0.01
multiple	0.01	frequent	0.01	chip	0.01	design	0.01
recognition	0.01	e	0.01	natural	0.01	engine	0.01
relations	0.01	page	0.01	cortex	0.01	service	0.01
library	0.01	gene	0.01	spike	0.01	social	0.01

? ? ?

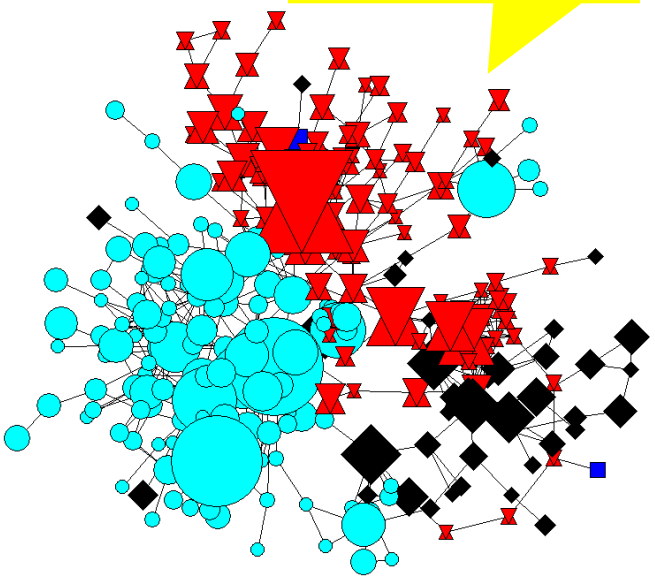


Topical Communities with NetPLSA

Topic 1		Topic 2		Topic 3		Topic 4	
retrieval	0.13	mining	0.11	neural	0.06	web	0.05
information	0.05	data	0.06	learning	0.02	services	0.03
document	0.03	discovery	0.03	networks	0.02	semantic	0.03
query	0.03	databases	0.02	recognition	0.02	services	0.03
text	0.03	rules	0.02	analog	0.01	peer	0.02
search	0.03	association	0.02	vlsi	0.01	ontologies	0.02
evaluation	0.02	patterns	0.02	neurons	0.01	rdf	0.02
user	0.02	frequent	0.01	gaussian	0.01	management	0.01
relevance	0.02	streams	0.01	network	0.01	ontology	0.01

Web

Coherent community assignment



Information Retrieval

Data mining

Machine learning

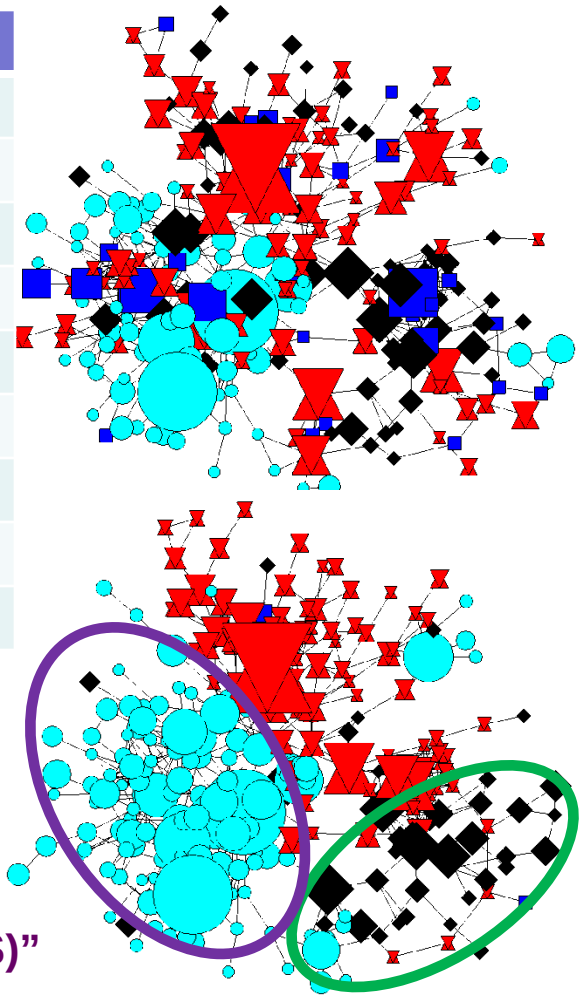
Coherent Topical Communities

NetPLSA	
neural	0.06
learning	0.02
networks	0.02
recognition	0.02
analog	0.01
vlsi	0.01
neurons	0.01
gaussian	0.01
network	0.01

PLSA	
visual	0.02
analog	0.02
neurons	0.02
vlsi	0.01
motion	0.01
chip	0.01
natural	0.01
cortex	0.01
spike	0.01

Semantics of community:

"machine learning (NIPS)"



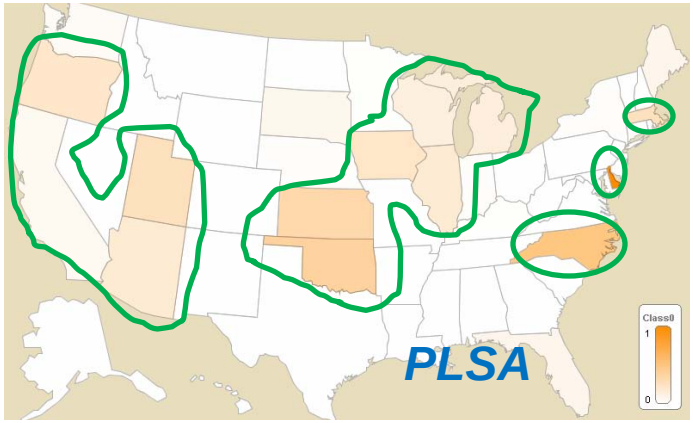
PLSA	
peer	0.02
patterns	0.01
mining	0.01
clusters	0.01
stream	0.01
frequent	0.01
e	0.01
page	0.01
gene	0.01

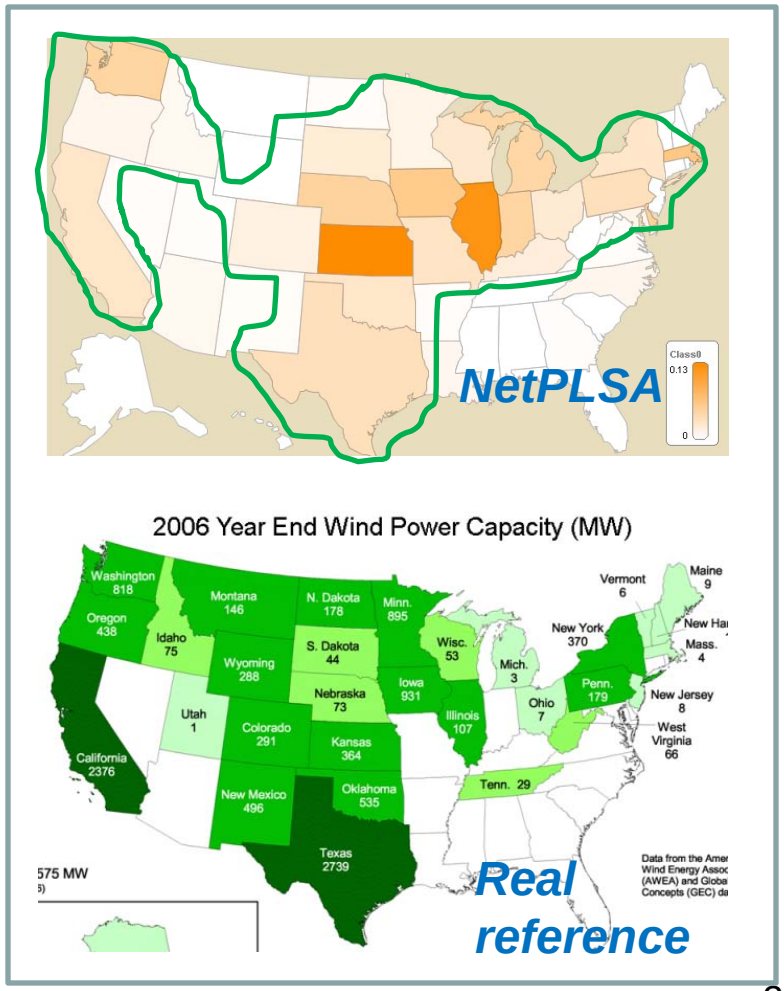
Semantics of community:

"Data Mining (KDD) "

NetPLSA	
mining	0.11
data	0.06
discovery	0.03
databases	0.02
rules	0.02
association	0.02
patterns	0.02
frequent	0.01
streams	0.01

Smoothed Topic Map





The Windy States

-Blog articles: "weather"

-US states network:

-Topic: "windy"

Summary

- Combine Topic modeling and network analysis
- A unified optimization framework
- NetPLSA = PLSA + Network Regularization
- Topical communities and topic map
- Future work:
  - Using other topic models (e.g., LDA)
  - More network properties (e.g., small world)
  - Topic evolution/spreading on network

Thanks!