

Topic Modeling with Network Regularization

Qiaozhu Mei, Deng Cai, Duo Zhang,
ChengXiang Zhai

University of Illinois at Urbana-Champaign

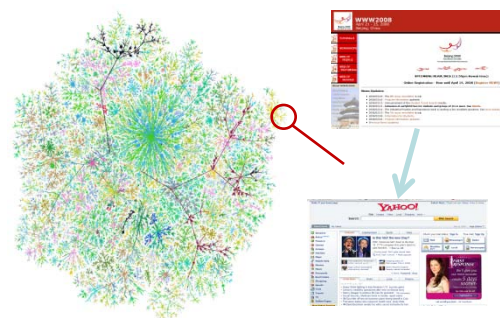
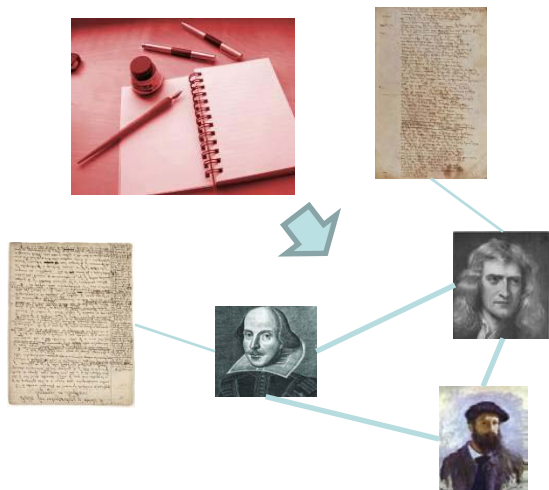
Outline

- 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

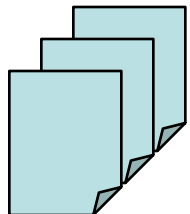
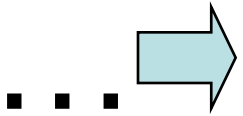
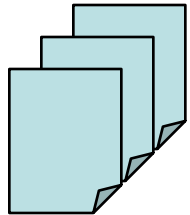


Web page + hyperlink structure

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

Probabilistic Topic Models for Text Mining

Text Collections



Probabilistic Topic Modeling

PLSA [Hofmann 99]
LDA [Blei et al. 03]
Author-Topic [Steyvers et al. 04]
Pachinko allocation [Li & McCallum 06]
CPLSA [Mei & Zhai 06]
CTM [Blei et al. 06]
 ■ ■ ■

Topic models
(Multinomial distributions)

term 0.16
relevance 0.08
weight 0.07
feedback 0.04
independ. 0.03
model 0.03
 ...

Subtopic discovery

Topical pattern analysis

Summarization

web 0.21
search 0.10
link 0.08
graph 0.05
 ...

Opinion comparison

...

Usually don't include network structure

Social Network Analysis

A Web Site as a Living Organism

Social and computer scientists are studying how social networking Web sites, like myspace.com, grow and change. They hope to learn why and how some online groups thrive and attract members while others stagnate and die out.

EACH CIRCLE represents one member. Larger circles are members who recruited more new members.

Old member
New member

EACH LINE represents a "friendship" between two people.

Old friendship
New friendship

Corresponds to someone recruiting a friend into the group.



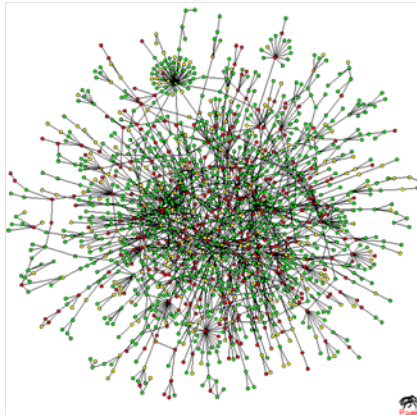
DEAD AREA
Part of the group with many existing members who are doing very little to actively recruit new people.

AREAS OF GROWTH
Part of the network where many new members are joining through their connections to existing members.

Source: Jon Kleinberg and Lars Backstrom, Cornell University

David Cozzani/The New York Times

- 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

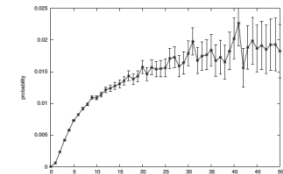
[Gruhl 04]; [Backstrom 06]

Ranking

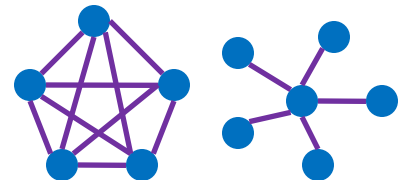
e.g., [Brin and Page 98];
[Kleinberg 98]

Structural patterns

e.g., [Yan 02]

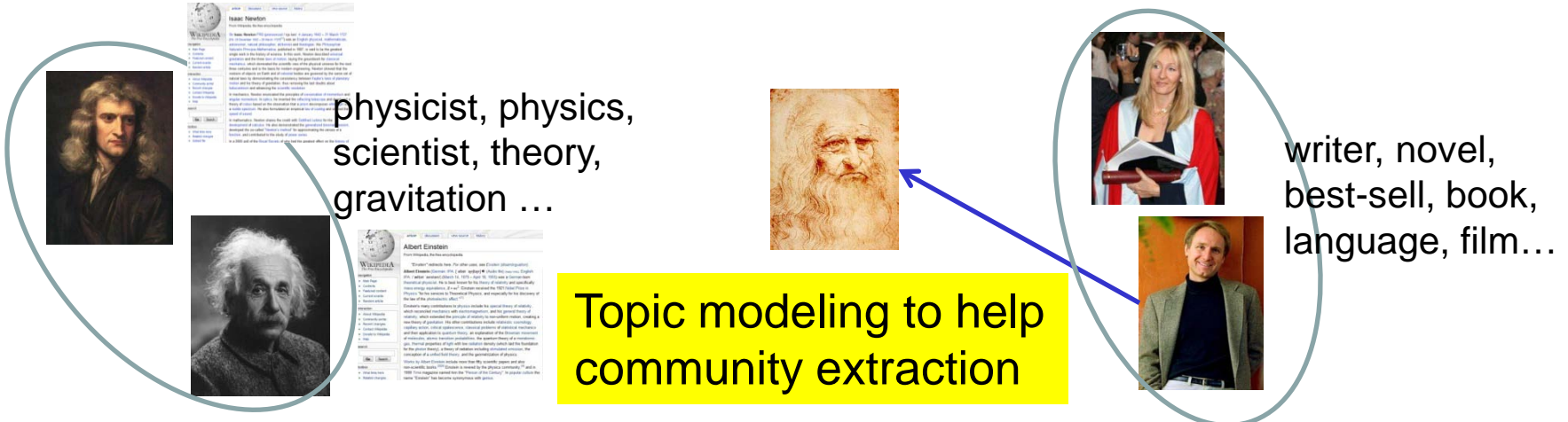


PageRank Update In Progress

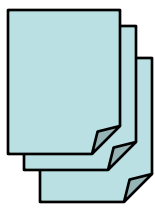
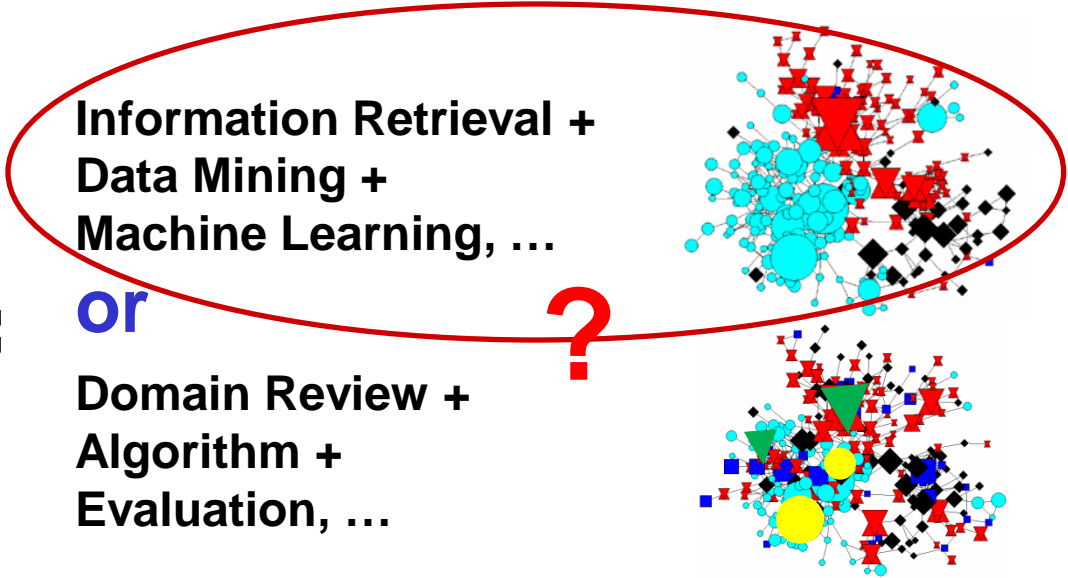


Usually don't model topics in text

Importance of Topic Modeling Plus Network Analysis



Network analysis to help topic extraction

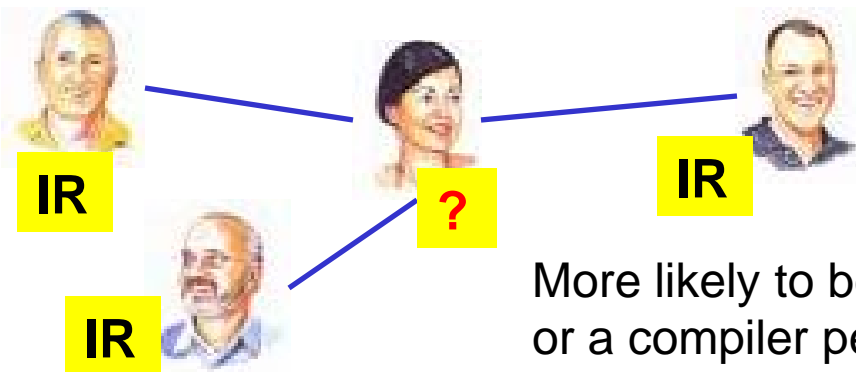


Computer Science Literature =

or Domain Review + Algorithm + Evaluation, ...

Intuitions

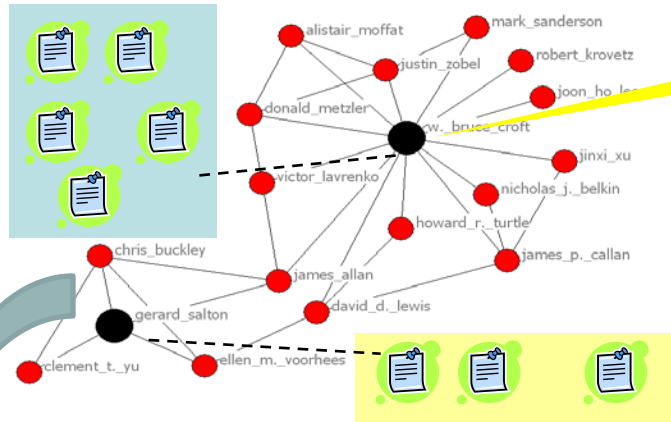
- People working on the same topic belong to the same “topical community”
- Good community: coherent topic + well connected
- A topic is semantically coherent if people working on this topic also collaborate a lot



Intuition: my topics are similar to my neighbors

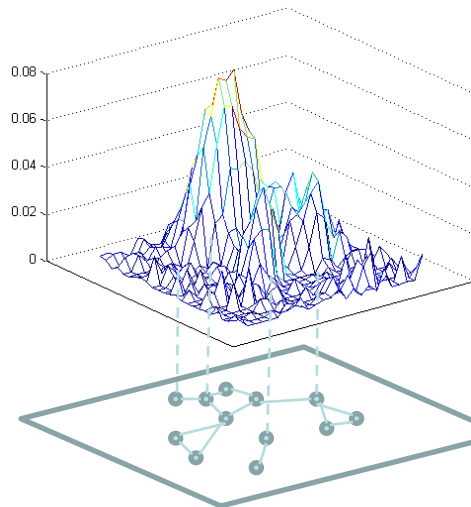
More likely to be an IR person or a compiler person?

Social Network Context for Topic Modeling

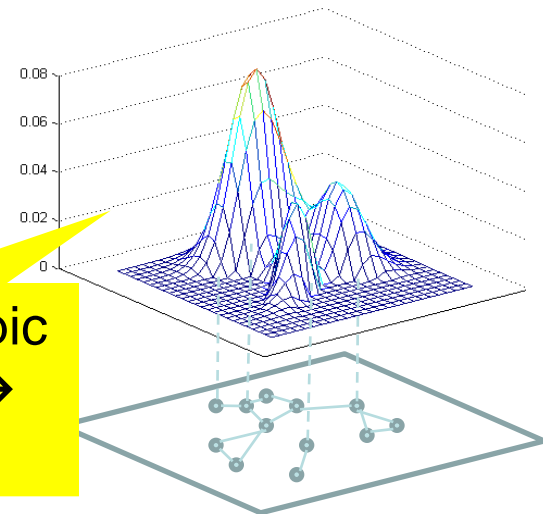


e.g. coauthor network

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



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



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):

$$O(\textit{Collection} \mid \textit{Model}) = \log(P(\textit{Collection} \mid \textit{Model}))$$

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

$$O(\textit{Collection}, \textit{Network} \mid \textit{Model})$$

$$= \log(P(\textit{Collection} \mid \textit{Model})) \oplus \textit{Regularizer}(\textit{Model}, \textit{Network})$$

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

PLSA : Probabilistic Latent Semantic Analysis (Hofmann '99)

- Generation Process:

c: the occurrences of word in doc

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

conditional probability => distribution

Topics $\theta_{1...k}$



Draw a word from
government 0.3
response 0.2..

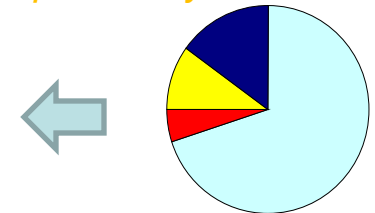


donate 0.1
relief 0.05
help 0.02 ..



city 0.2
new 0.1
orleans 0.05 ..

Criticism of government response to the hurricane primarily consisted of **criticism** of its **response** to ... The total **shut-in oil production** from the Gulf of Mexico ... approximately 24% of the **annual production** and the **shut-in gas production** ... Over seventy countries **pledged monetary donations** or other **assistance**. ...



$P(\theta_i/d)$

Can be context-sensitive

Choose a topic

NetPLSA

- Basic Assumption: Neighbors have similar topic distribution

topic distribution of a document

PLSA

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

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

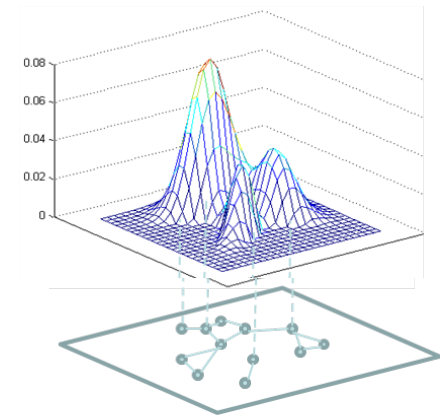
tradeoff

importance (weight) of an edge

difference of topic distribution

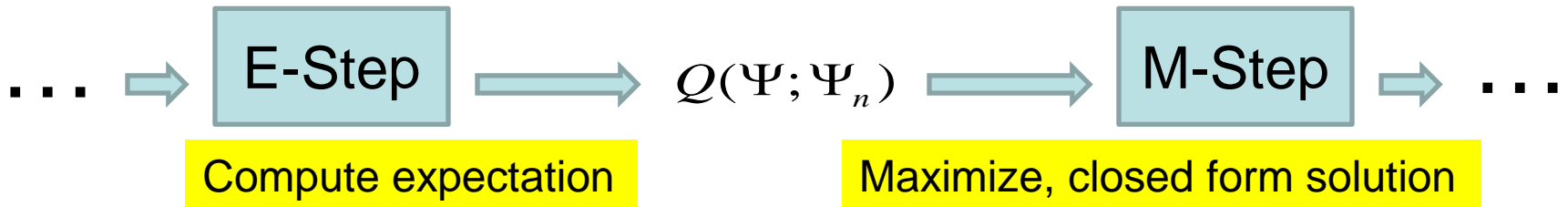
Graph Harmonic Regularizer, Generalization of [Zhu '03],

$$= \frac{1}{2} \sum_{j=1 \dots k} f_j^T \Delta f_j, \text{ where } f_{j,u} = p(\theta_j | u)$$

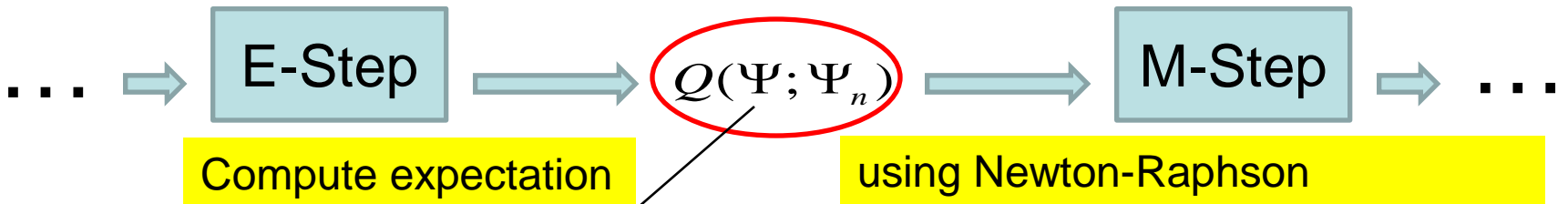


Parameter Estimation

- PLSA: EM algorithm



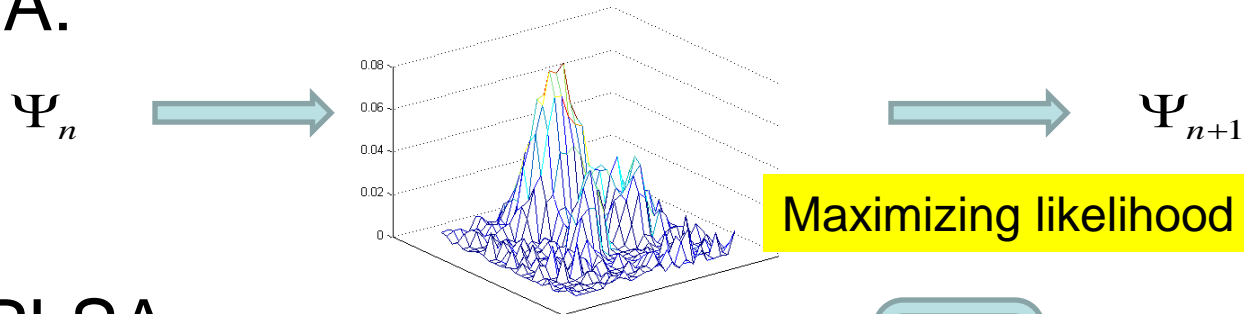
- NetPLSA: Generalized EM Algorithm



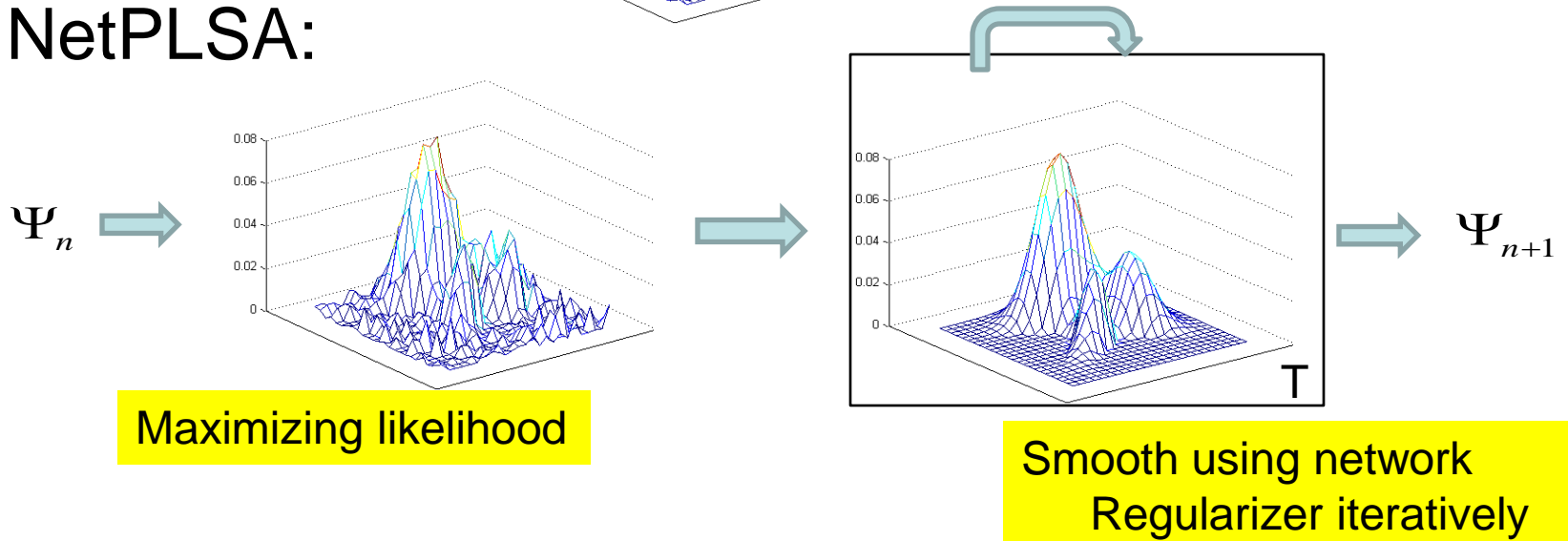
Regularized complete likelihood

How it Works in M Step

- PLSA:

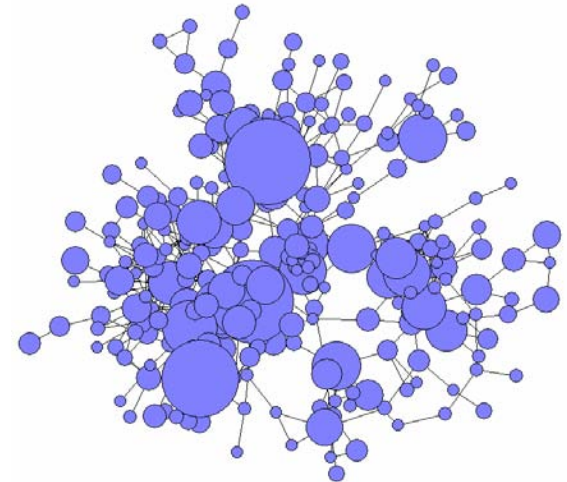


- NetPLSA:



Experiments

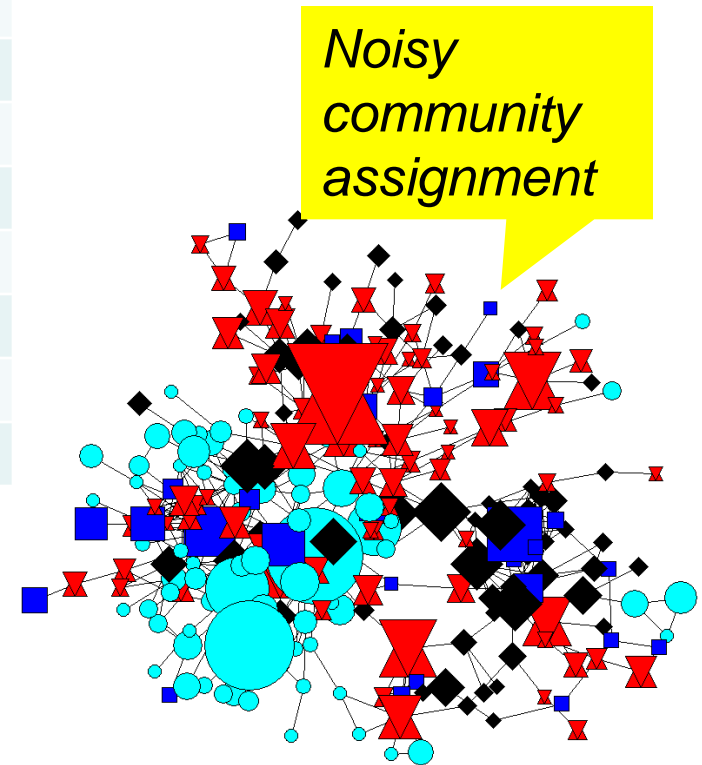
- Bibliography data and coauthor networks
 - DBLP: text = titles; network = coauthors
 - Four conferences (expect 4 topics): SIGIR, KDD, NIPS, WWW
- Blog articles and Geographic network
 - Blogs from spaces.live.com containing topical words, e.g. “weather”
 - Network: US states (adjacent states)



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

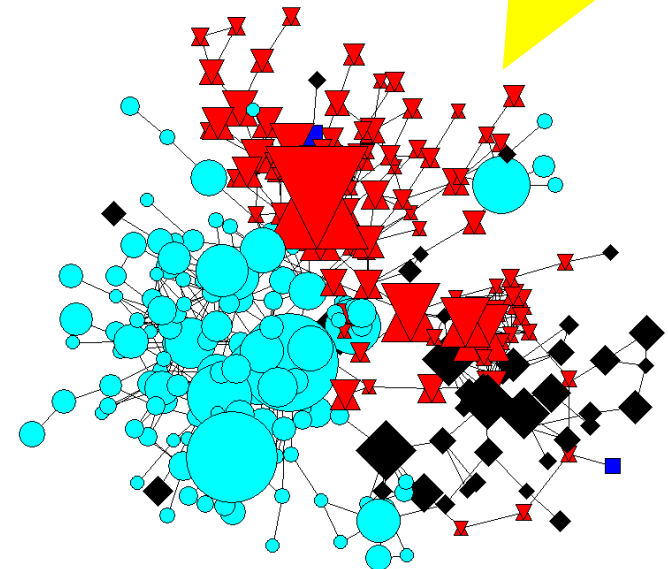
Web

Coherent community assignment

Information Retrieval

Data mining

Machine learning

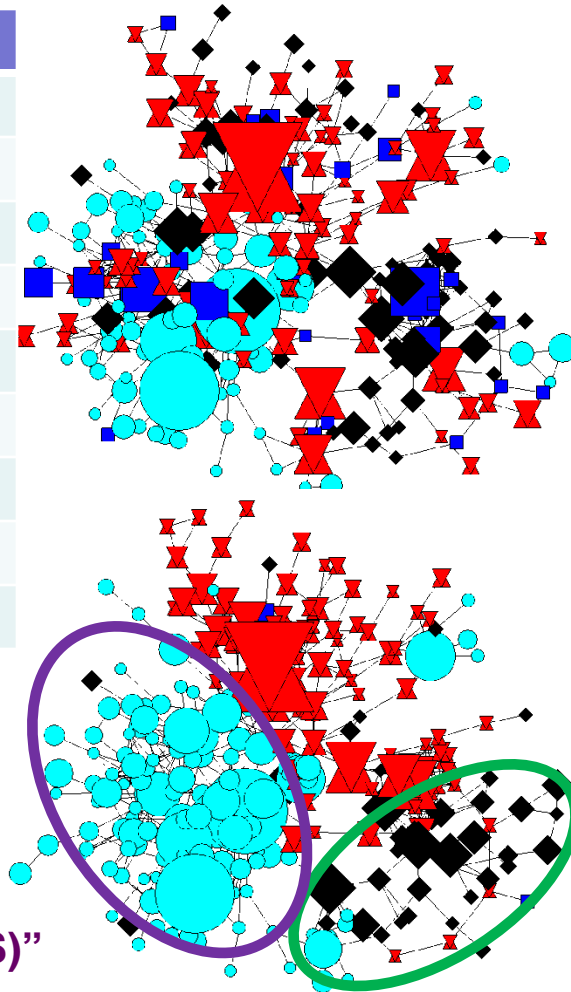


Coherent Topical Communities

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)”

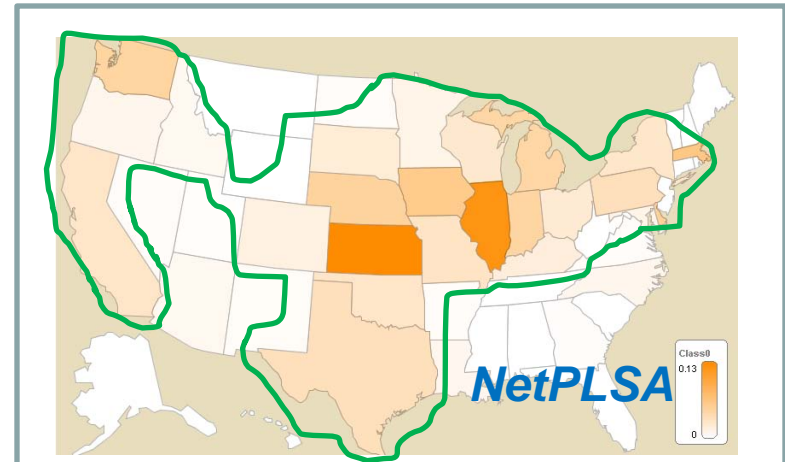
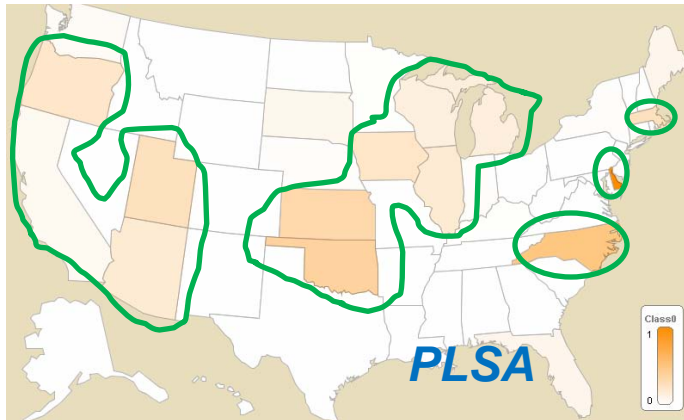


PLSA	
peer	0.02
patterns	0.01
mining	0.01
clusters	0.01
stream	0.01
frequent	0.01
e	0.01
page	0.01
gene	0.01

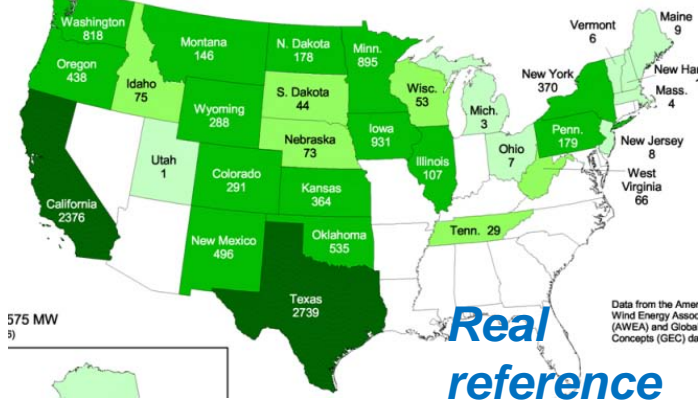
Semantics of community:
“Data Mining (KDD)”

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



2006 Year End Wind Power Capacity (MW)



- 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!