ICS – SOCIAL NETWORK DISCOVERY
Presented by - Karan Kurani and Jason Marcell
(Some slides adapted from presentation on 12th November)
OVERVIEW

- Goal
- Datasets
- Software Engineering
- Latent Dirichlet Allocation
- Methodology
- Results
- Future Work
GOAL

- Find people who are doing Comp Sust. But who are not aware about it or we don’t know about them.

- Techniques –
  - Citation Network Analysis (Not implemented yet)
  - Similarity Measure
  - Combination of both.
DATASETS

- CS Based - DBLP, arnetminer.org, CiteSeerX.
- Multidisciplinary – BASE, Bioone, ChemSeerX, Crossref for citation.
- Currently Used –
SOFTWARE ENGINEERING PRACTICES

Revision Control
Logging
Unit Testing
Object-Relational Mapping
Integrated Development Environment
**APPROACHES**

- **DBLP Stats:**
  - Total docs: 1632441
  - With abstract text: 653507
  - With references: 316559

- Possible approaches included –
  - LSA, pLSA and LDA.
  - All of them make a bag of words model.
LATENT DIRICHLET ALLOCATION

computer chemistry cortex orbit infection
methods synthesis stimulus dust immune
number oxidation fig jupiter aids
two reaction vision line infected
principle product neuron system viral
design organic recordings solar cells
access conditions visual gas vaccine
processing cluster stimuli atmospheric antibodies
advantage molecule recorded mars hiv
important studies motor field parasite

FIGURE 1. Five topics from a 50-topic LDA model fit to Science from 1980–2002.

*From the review paper “Topic Models” - David M. Blei, Princeton University. John D. Lafferty, Carnegie Mellon University
APPLICATIONS OF LDA

Images (Fei-Fei and Perona, 2005; Russell et al., 2006; Blei and Jordan, 2003; Barnard et al., 2003),

Population genetics data (Pritchard et al., 2000),

Survey data (Erosheva et al., 2007),

Social networks data (Airoldi et al., 2007).
LDA WITH MAHOUT

1. DBLP Data Set
2. CompSust Keyword Filter
3. Stop Words Filter
4. MAHOUT LDA
   - Extract corpus and seed paper topic distributions
   - Squared Euclidean Distance
   - Cosine Distance
   - Symmetric KL-divergence distance
LDA USING LINGPIPE

1. DBLP+ Citation Dataset → Filtering
2. Papers with Abstract → Stemming and Stop Word Removal
3. Tokens → Tokenizer
4. Seed Paper Set → Stemming and Stop Word Removal
5. Tokenization → LDA Model
6. Seed Paper Topic Distribution
7. KL Divergence Calculator
8. Sorted List of papers according to KL Divergence
Evolving results set can be browsed on the web:
http://www.cs.cornell.edu/~kiyan/compsustsn/
RESULTS AFTER A MONTH

- Noisy but Encouraging (Most of the results are recent (2006-2010.))
- Reasons -
  - Many false positives because of alternate uses of keywords.
  - Over fitting because of sub optimal parameters for LDA.
BUILDING ON LDA – SOME MORE MODELS

Correlated Topic Models

Dynamic Topic Models
NEXT STEPS

- Add additional data sources.
- Customized web crawler.
- Incorporate network analysis (Author – topic model, Link-LDA)
THANKS!

Questions are guaranteed in life; Answers aren't.