Finding Semantic “Fingerprints” in Documents

Lashon B. Booker, Ph.D.

E540

booker@mitre.org

ASIAS Technology and Tools Symposium

July 28, 2009

MITRE

Introduction

- **Text documents, messages and information exchanges can provide important insights about the structure and dynamics of social and information networks**
  - **Computational approaches to content and authorship analysis usually focus on structural characteristics and linguistics patterns in the text**
  - **The ideas and semantic content conveyed in the text are also important, but how can this analysis be done computationally?**
- **Many factors influence the ideas present in a document**
  - **In particular, factors related to social identity - such as the author's age or gender, ideology, beliefs, etc. – play an important role in communication behaviors.**
  - **If the attributes of these factors could be teased out of a document, they might provide a valuable "fingerprint" facilitating author analysis**
- **We describe preliminary experiments on a computational approach to extracting identity group fingerprints from text**

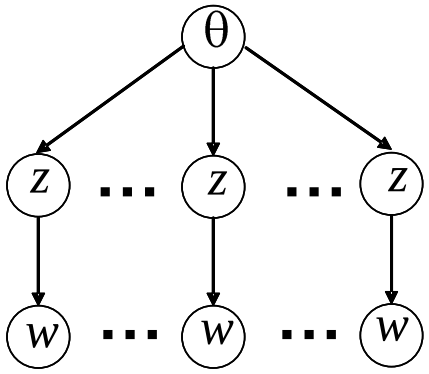
Background

- Our MITRE research project has developed approaches to modeling social phenomena such as group formation and group recruitment.
- Traditional models rely on social network analysis and focus on the behavior of individuals
- Analyzing the influence of group-level attributes may offer a way around the need for data about individuals and their relationships
  - Individuals can be drawn to, and influenced by, the attributes of a group and not specifically to any of its members
  - Changes in individual behavior are influenced, in part, by the ideas and beliefs that bind individuals to social identity groups and bias individual behavior
- A key research objective was to develop computational techniques for extracting group-level attributes from group artifacts (e.g. document collections)
  - The techniques developed so far have shown promise as an approach to document classification, authorship analysis, and sentiment analysis

Computing Identity Group Attributes

- **Case study: the formation of collaborative networks associated with scientific publications in a field of study**
  - **Publication venues are visible manifestations of various social identity groups**
  - **The topics and keywords used in documents published by group members provide information about relevant scientific concepts, issues, and positions on those issues (i.e. attributes)**
  - **Topic analysis techniques are available to extract topic information from a corpus of documents**
- **Question: do groups express a "voice" in these document collections that is independent of particular authors or documents? (i.e. group-level versus document-level attributes)**

Topic Analysis - Latent Dirichlet Allocation



Document "Gist"

Topics $z = (z_1, \ldots, z_T)$

Words $w = (w_1, \ldots, w_N)$

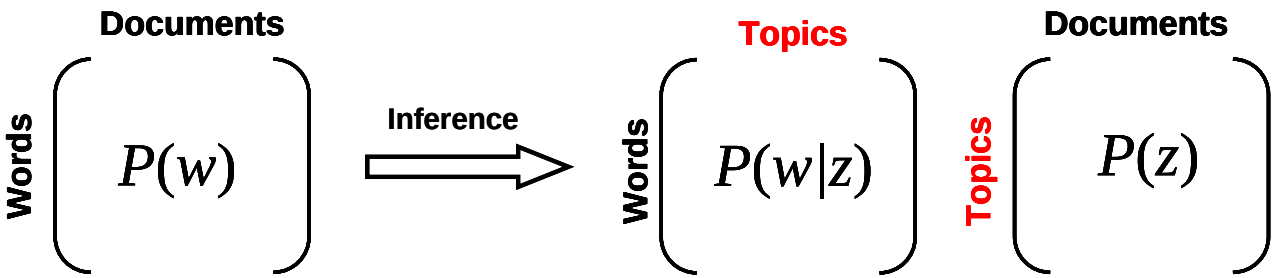
$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j)$$

- **Latent Dirichlet Allocation* (LDA) is a generative probabilistic model for documents**
  - **Assumes a latent structure in the corpus consisting of a set of "topics"**
  - **Every document is a mixture of topics and every topic is a mixture of words.**
- **The gist of a document is represented using a probability distribution over T topics**
  - **Uses a "bag of words" representation for each document (i.e., word order is ignored)**
  - **The mixture proportions are modeled using multinomial distributions, and Dirichlet priors model how the proportions vary.**
  - **Each word is chosen from a single topic (potentially different for each word)**

*Blei, D., Ng, A. and Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research,* **3**:993-1022

MITRE

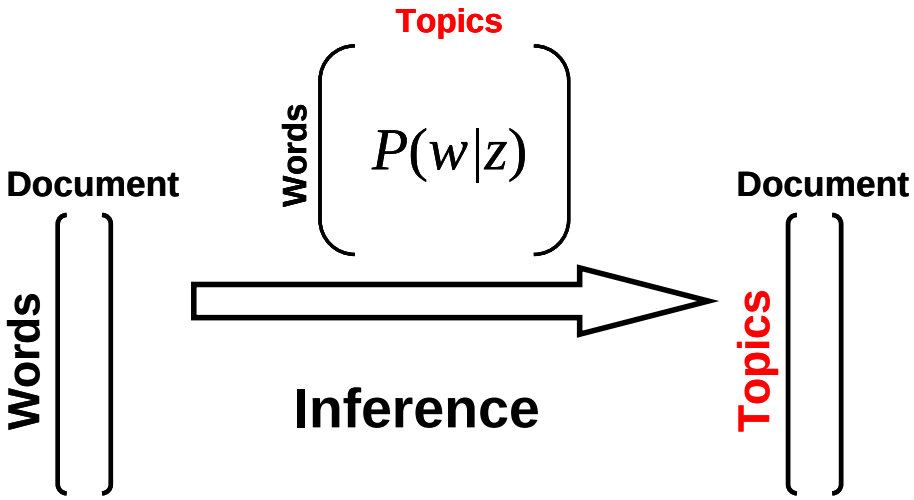
Extracting Topics From a Corpus



- Statistical inference methods can be used to infer the LDA model parameters for generating the corpus. We use Gibbs sampling*.
  - The word-document co-occurrence matrix is the only input needed from the corpus.
  - The inference method tries to strike a balance between having few prominent topics per document and few prominent words per topic
  - The outputs are the set of topics responsible for generating a document collection, and the topic distribution for each document.
- The set of topics generated by this procedure is not unique.

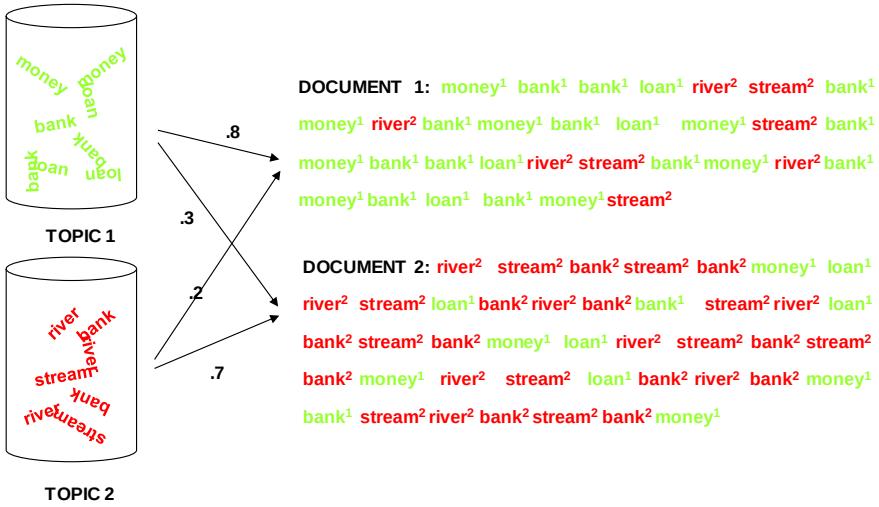
*Griffiths, T. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Science, 101*, 5228-5235.

Analyzing Topics in a New Document

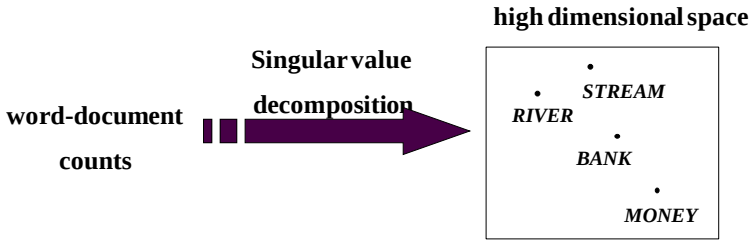


- The most rigorous approach is to add the new document to the existing corpus and rerun the algorithm
- A more computationally efficient approach is to run the inference algorithm on the new document exclusively .
  - Re-use the topic definitions learned previously from the corpus by keeping the topic-word distribution fixed.
  - The set of topic proportions inferred for the new document can differ significantly from any document in the original corpus
  - Assignments of topics to words are generated in this process as well

LDA vs LSA (Griffiths, Steyvers &Tenenbaum ,2007)



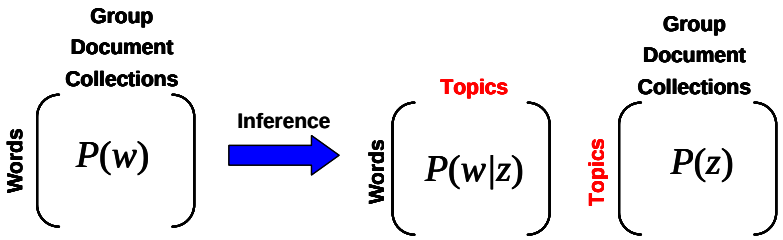
- **Latent Dirichlet Allocation (LDA)**
  - **Generative model with meaningful topics and extendible structure**
  - **Polysemy: different word meanings can be represented by different topics**
  - **Synonymy: similar words have comparable probabilities in the same topics**
  - **Outperforms LSA in predicting word association and explaining various aspects of human semantic representations**



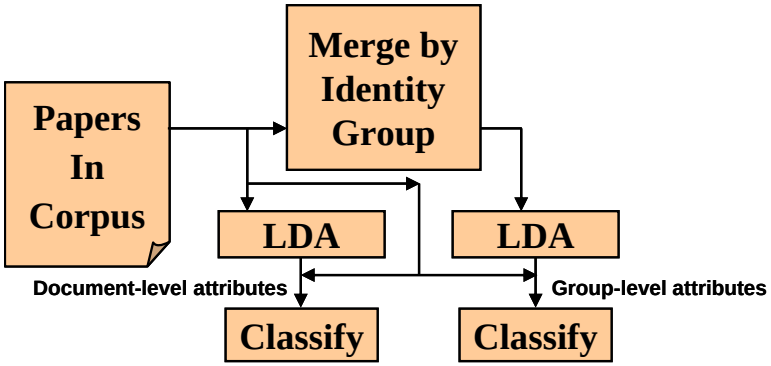
- **Latent Semantic Analysis (LSA)**
  - **Not generative, difficult to extend**
  - **Each word is only a single point in semantic space**
  - **Similarity measured by cosine of angle between word vectors**

Griffiths, T., Steyvers, M., and Tenenbaum, J. (2007). Topics in semantic representation. *Psychological Review, 114(2)*, 211-244.

Group-level vs Document-level Attributes



Topic analysis at the "group level"



Two alternative representations for classifying documents

- We propose that group-level topics might be teased out of the corpus by focusing on the word co-occurrences within each sub-collection of group documents
  - Provide aggregated co-occurrence data to LDA by merging group documents
  - The topics computed will thereby be attributes of the group, not attributes associated with the individual documents
- Group-level attributes can also be used as features for individual documents
  - If these attributes really capture the group "voice" in the documents, they should perform better on classification tasks

A Document Classification Experiment

- **Used the document collection from the Information Visualization 2004 Contest Dataset***
  - **Includes 614 papers (no full text) published by 1036 unique authors between 1974 and 2004 in the field of Information Visualization.**
  - **Selected the 429 documents in the dataset that have abstracts. The title, abstract and keywords constitute the "bag of words" for each document**
  - **Preprocessing involved removing punctuation, single characters, 2-character words, words on a linguistic "stop list", and words appearing less than 5 times in the corpus**
- **Assumed each source "venue" having 10 or more publications specifies an identity group in this field**

IEEE Symposium on Information Visualization

IEEE Visualization

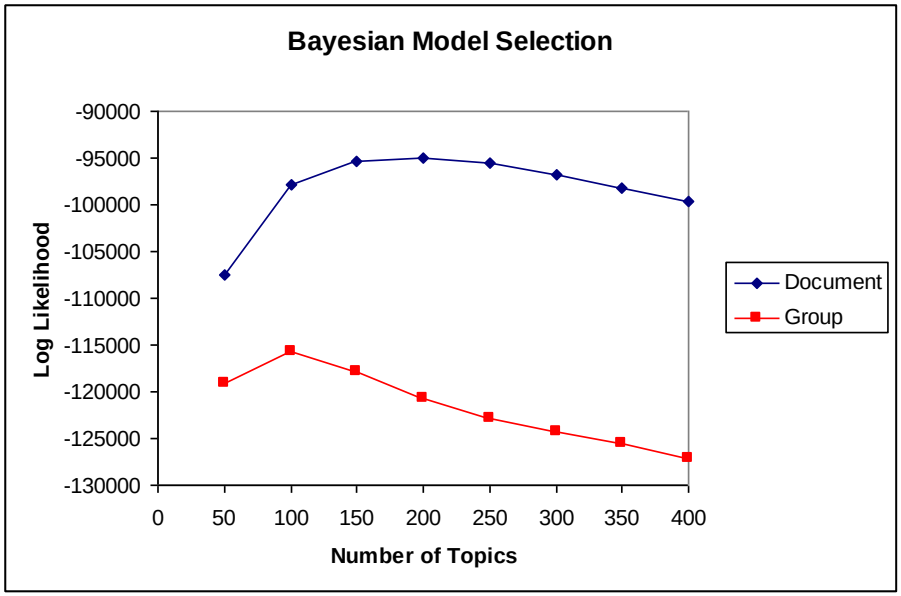
Lecture Notes in Computer Science

Conference on Human Factors in Computing Systems

etc.

***K. Borner, et al. (2005) "Studying the Emerging Global Brain: Analyzing and Visualizing the Impact of Co-Authorship Teams." Complexity, Vol. 10, No. 4.**

Choosing the Number of Topics



- **The optimal number of topics that fits the data well without overfitting can be determined using a Bayesian method for solving model selection problems**
  - **Find the value of the model parameter T (number of topics) that maximizes the likelihood of the observed data (the corpus *w*) (i.e., compute P(*w* | *T*))**
  - **Using sample values generated by the LDA algorithm, the likelihood for each parameterized model can be estimated analytically**
- **For the InfoVis corpus, the model with 100 topics had the highest likelihood at the group level and the 200 topic model was best at the document level**

Classification Results

Identity Group	# Samples	Document level Accuracy	Group level Accuracy
ACM CSUR	210	88.91%	96.91%
Advanced Visual Int	100	74.30%	98.06%
CACM	100	80.11%	91.77%
CGIT	120	82.01%	100.00%
IEEE Comp Graphics	130	80.87%	94.40%
IEEE Symp on InfoViz	1520	77.49%	87.34%
IEEE Transactions	130	71.81%	87.60%
IEEE Visualization	320	80.03%	90.94%
LNCS	220	95.40%	97.12%
SIGCHI	210	87.35%	90.06%
UIST	170	93.68%	90.73%
Other	1060	75.31%	83.89%
Overall	**4290**	**79.74%**	**88.70%**

- **Ten examples were generated for each representation (document-level and group-level) of each document.**
- **A support vector machine classification algorithm* was used to solve the multi-class classification problem.**
  - **Results are averaged over 10 independent runs. On each run 75% of the examples were used for training and the remaining 25% were used for testing.**
- **The group-level features produced substantially better overall accuracy than the document-level features on test data.**

*T. Joachims (1999). Making large-scale SVM learning practical. In B. Schölkopf, C. Burges and A. Smola (Eds.), Advances in Kernel Methods – Support Vector Learning. MIT Press.

Application to Blog Authorship Analysis

- **The Blog Authorship Corpus* was constructed using blogs collected from blogger.com in August 2004**
  - **The corpus consists of 19,320 blogs containing 681,288 posts and over 140 million words**
  - **It includes 3 age categories, each with an equalized gender distribution**
  - **Each blog was represented using a set of carefully-chosen vocabulary features (502 style-related and 1000 content-related)**
  - **Classification models derived from these features predict the correct age with an accuracy of 76.2% and the correct gender with an accuracy of 80.1%.**
- **Topic analysis can be used to automatically generate features that may be helpful for author profiling.**
  - **Classification models (100 topics) derived from group-level features predict the correct age with an accuracy of 75.2% and the correct gender with an accuracy of 76.3%.**

*Schler, J., Koppel, M., Argamon, S., Pennebaker, J.: Effects of Age and Gender on Blogging. In: Computational Approaches to Analyzing Weblogs: Papers from the 2006 AAAI Spring Symposium. AAAI Press, Menlo Park (2006)

Application to Sentiment Categorization

- **We revisit a previous study* on using the words in a critic's movie review to predict how that critic rates the movie.**
  - **The Scale Dataset was constructed from movie reviews written by 4 authors**
    - **Reviews were preprocessed to remove explicit rating indicators and objective sentences**
    - **Class labels were derived by normalizing each reviewer's ratings to fit a three-category rating system (negative, middling, or positive)**
    - **The corpus consists of 5006 reviews containing 16,244 words**
  - **Classification models using words as features predict the correct rating with an accuracy of about 67%.**
- **Classification models derived from group-level LDA models of the rating categories (50 topics) predict the correct rating with an accuracy of 78.39%**

*Pang, B. and Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the ACL* (pp. 115-124), Ann Arbor, June 2005, Association for Computational Linguistics.

MITRE

Prospects for ASIAS Applications

- **Preliminary document classification experiments have been conducted using samples of ASRS safety reports**
  - **Using a collection of 199 reports describing flights in 2 categories (loss of control – true or false), the LDA model achieved 91.8% accuracy, with a 2.42% false positive rate**
  - **Using a collection of 6138 reports covering 15 categories (maintenance, weather, passenger, etc.), the LDA model achieved an overall accuracy of 76%.**
  - **Theses results compared favorably with (i.e., were generally better than) results achieved on this data using other methods**
- **Only the "vanilla" version of LDA was used. Better classification results could probably be obtained using methods to account for unbalanced data sets, preprocessing to identify meaningful phrases, tuning the SVM parameters, etc.**
- **It might also be helpful to use LDA semantic analysis to get a better understanding of these data sets (examine the prominent topics, find clusters in the set of documents, etc.)**

Summary

- We have shown how identity group "fingerprints" can be extracted from a document collection by applying topic analysis methods in a novel way
  - Empirical results suggest that group attributes provide better predictions of document content than attributes derived from the individual documents
  - Preliminary investigations suggest that similar results can be obtained with document collections associated with many kinds of identity groups and classification problems.
- Experiments with blog data show that this document classification method can also be effective for forensic authorship analysis tasks
  - Besides providing good classification accuracy, it has the added benefit of automatically inferring a useful set of features
  - This capability could be a useful supplement to forensic methods that incorporate other techniques such as linguistic analysis and behavioral profiling
- Preliminary results with a sentiment analysis task are also promising

Related Text Mining Applications of LDA

- **Extract relationships between topics and named entities.**
  - **D. Newman, C. Chemudugunta,P. Smyth and M. Steyvers, "Analyzing Entities and Topics in News Articles UsingStatistical Topic Models." *LNCS – Proceedings of the IEEE Conference on Intelligence and Security Informatics*, pp93-104,Springer-Verlag, 2006.**
- **Analyze temporal dynamics of the content in a corpus, identifying trends in topics over time.**
  - **T. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Science, vol. 101, pp. 5228–5235, 2004.***
  - **X. Wang, W. Li, and A. McCallum. "A Continuous-Time Model of Topic Co-occurrence Trends". Proceedings of the AAAI Workshop on Event Detection, 2006.**
- **Extract semantic content from annotated data (e.g. images and their captions, papers and bibliographies, genes and their functions).**
  - **D. Blei and M. Jordan, "Modeling annotated data," *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 127–134, New York: ACM, 2003.***
  - **V. Jain, E. Learned-Miller, and A. McCallum. People-LDA:Anchoring topics to people using face recognition. *Proceedings of the IEEE International Conference on Computer Vision, pp. 1–8, 2007.***
- **Use topic analysis to detect unusual papers by a specific author or attribute authorship to words in a jointly authored document.**
  - **M. Steyvers, P. Smyth, and T. Griffiths. "Probabilistic author-topic models for information discovery," *Proceedings of the 10th ACM SIGKDD International conference on knowledge discovery and data mining, pp. 306–315, New York: ACM, 2004.***

Backup Slides

Using Topics for Document Classification

- One outcome of estimating the parameters of an LDA model is that documents are represented with a fixed set of real-valued features
  - The topic distribution can be interpreted as a set of normalized feature weights with the topics as the features
  - These document descriptions are ideally suited to serve as example instances in a document classification problem
- Results in the literature suggest that features induced by an LDA model are as effective as using individuals words as features for classification, but with a big advantage in dimensionality reduction
- Document classification provides a good way to assess how well a set of attributes characterizes class membership
  - How might we compute group-level attributes useful for classifying identity groups?

DiagnosticTopics

IEEE Computer Graphics and Applications

Top 3 document-level topics

Topic 36 (0.6249)	Topic 118 (0.5912)	Topic 3 (0.4379)
diagram (0.314)	match (0.156)	color (0.219)
theory (0.117)	system (0.094)	attributes (0.107)
representing (0.080)	tennis (0.086)	identify (0.077)
cognitive (0.051)	varying (0.070)	help (0.047)
effectiveness (0.044)	map (0.055)	produce (0.041)
perceptual (0.044)	competition (0.055)	picture (0.041)
processes (0.044)	consist (0.039)	similar (0.036)
external (0.029)	explore (0.039)	linearized (0.030)
range (0.022)	primary (0.039)	medical (0.030)
solving (0.022)	top (0.031)	show (0.030)

Top 3 group-level topics

Topic 38 (0.9297)	Topic 79 (0.6848)	Topic 17 (0.4373)
diagram (0.165)	knowledge (0.097)	nodes (0.155)
match (0.108)	color (0.073)	element (0.094)
scale (0.057)	visage (0.065)	map (0.078)
plant (0.044)	relational (0.065)	link (0.073)
medical (0.038)	visual (0.057)	quickly (0.065)
discusses (0.038)	simultaneously (0.045)	operation (0.061)
node (0.038)	distortion (0.045)	easily (0.057)
routine (0.038)	types (0.032)	describe (0.049)
linearized (0.025)	test (0.028)	tables (0.033)
processes (0.025)	generates (0.024)	competition (0.029)

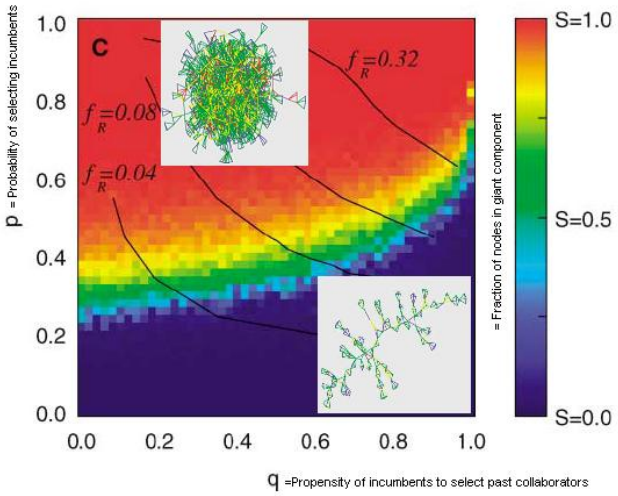
- If $\theta_{i,j}$ is the probability of topic j in mega-document i, then the diagnostic value of topic j for class i is given by the ratio $\theta_{i,j} / \sum_k \theta_{k,j}$ where the sum is over all classes k
- Group-level topics tend to be more diagnostic than document-level topics.
- Document-level topics seem to distinguish groups by identifying themes and concepts expressed in particular documents
- Topic 38 captures the concept “interesting applications” - a group-level theme

Probability of randomly chosen word = 1/1405 = 0.0007

MITRE

Formation of an “Invisible College”

The emergence of large connected communities of practitioners in an “Invisible College” can be described as a phase change.

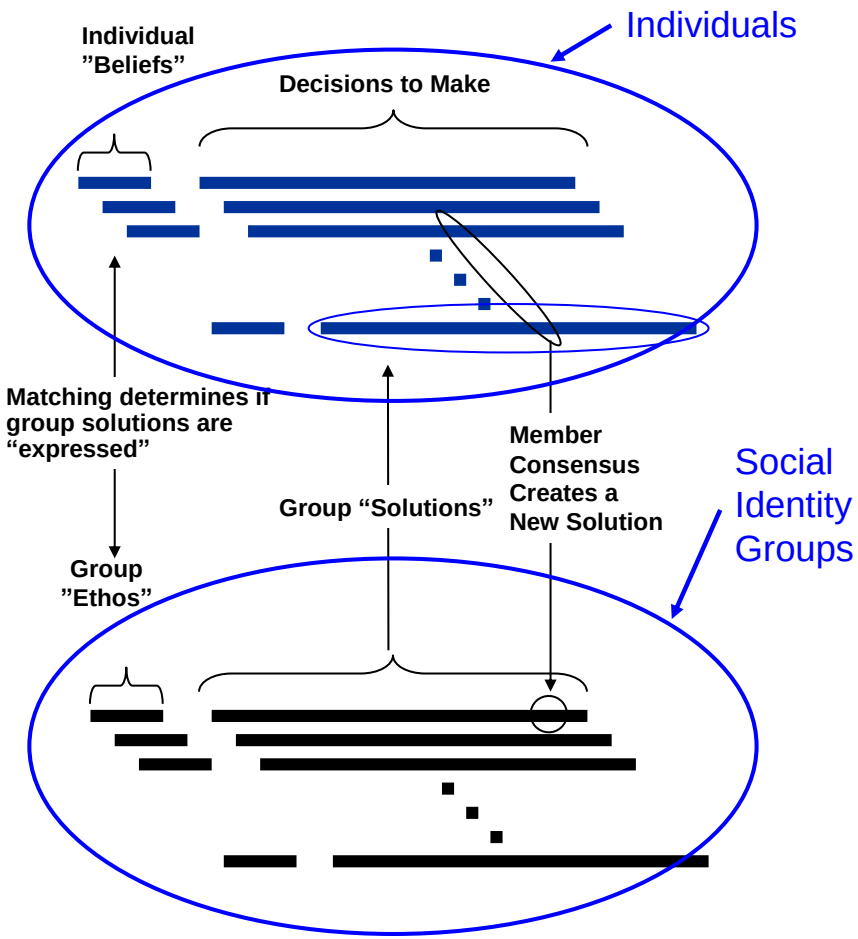


- **Team assembly models show phase-changes in the growth of collaborative networks**

  Guimera, et al. “Team assembly mechanisms determine collaboration network structure and team performance”, Science, Vol. 308, 697-702 2005.

- **In this figure, note the narrow range of *p* over which the system changes from one characterized as a large number of small clusters to one characterized by a single large cluster**

Modeling Framework



- When confronted with a problem to solve, individuals choose from a variety of social identity groups that offer solutions.
- Identity groups provide both clarity and a simplified approach to making complex decisions (only a small number of groups are "expressed").
- Changes in the way individual beliefs align with the group "ethos" may signal a phase change.