Topic Model Diagnostics: Assessing Domain Relevance via Topical Alignment

(Supplementary Materials)

Jason Chuang JCCHUANG@CS.STANFORD.EDU
Stanford University, 353 Serra Mall, Stanford, CA 94305 USA

Sonal Gupta SONAL@CS.STANFORD.EDU
Stanford University, 353 Serra Mall, Stanford, CA 94305 USA

Christopher D. Manning MANNING@CS.STANFORD.EDU
Stanford University, 353 Serra Mall, Stanford, CA 94305 USA

Jeffrey Heer JHEER@CS.STANFORD.EDU
Stanford University, 353 Serra Mall, Stanford, CA 94305 USA

1. Graphs

Supplementary Figure 2 shows an enlarged version of Figure 1 in the main paper with additional details in the caption.

Supplementary Figure 3 shows additional data points for Figure 7 in the main paper.

2. Expert-Authored Concepts in Information Visualization

We conducted a survey asking ten experienced information visualization (InfoVis) researchers to identify what they consider to be *significant and coherent areas of research* in their field. Participants were asked to label each area, and describe it with lists of exemplary terms and documents.

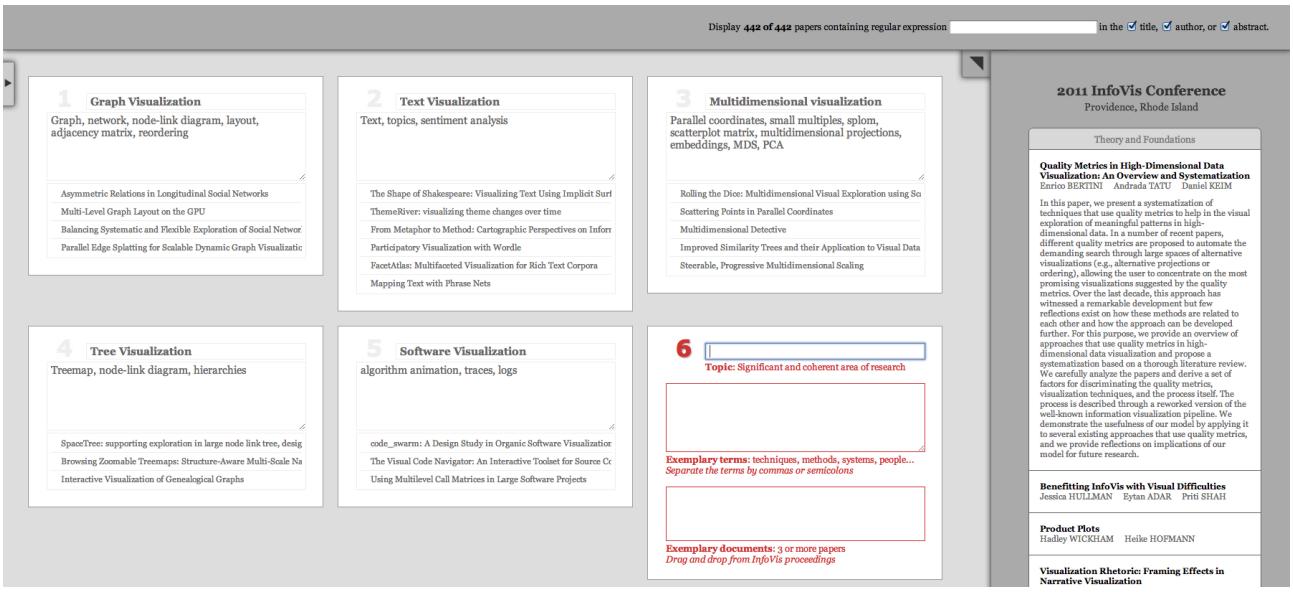
We focused on InfoVis research due to relevance, scope and familiarity. Analysis of academic publications is one of the common real-world uses of topic modeling (Griffiths & Steyvers, 2004). Our familiarity with the InfoVis community allowed us to contact experts capable of exhaustively enumerating its research areas. InfoVis has a single primary conference, simplifying the construction and analysis of its publications.

Survey recruitment was by invitation only. We contacted 23 researchers (12 past chairs of the IEEE Information Visualization Conference, six faculty members, two senior industry researchers, and three PhD students within a year of graduation) on a rolling basis over four months from March to June 2012. We sent out 14 surveys, and received ten completed results from four past chairs, two faculty members, one industry researcher, and three PhD students. We initially limited our survey to only past conference chairs, and expanded our criteria to established researchers to enable greater participation.

2.1. Survey Design

We asked participants to describe topics using labels, terms, and documents that they would use as if they were *communicating with a peer*. Representative terms should *exemplify a topic and differentiate the topic from other areas of research*. Terms could be any notable techniques, methods, systems, or people; multiword phrases were allowed. Representative documents *exemplify the core contributions of a topic*. Pilot studies suggested that citing a paper using freeform text is time consuming, disruptive to the recall process, and prone to errors. In response, we limited the representative papers to those published at IEEE Information Visualization Conference, and provided a drag-and-drop interface for associating a paper with a topic. We requested that participants enter ten or more terms and three or more papers per topic, though fewer responses were permissible. We asked participants to complete the survey in a single session if possible.

Conducted using a single webpage (Supplementary Figure 1), we designed the survey to (1) elicit expert responses with minimal bias, (2) support recall, (3)



Supplementary Figure 1: Survey user interface: Participants were provided with blank boxes in a single webpage, and asked to identify all *coherent and significant* areas of research in information visualization, in a manner as if communicating with a peer. An optional panel on the right shows 17 years of IEEE Information Visualization Conference proceedings.

enable accurate data collection, and (4) balance between maximizing the value of available expert time and preventing participant exhaustion.

To avoid artificially limiting what they consider to be the scope of InfoVis, the participants were instructed to consider work published anywhere when creating the research topics. Participants were provided with multiple blank boxes, and asked to enumerate all areas they consider to be significant. The webpage contained twenty boxes by default, but subjects could add additional boxes if desired.

In pilot studies, the single most prominent issue was recall. Exhaustively identifying all concepts in a domain purely from memory can be difficult. In response, we added a panel on the right that contains a list of all 442 papers published at the IEEE Information Visualization Conference (1995 to 2011), grouped by year. As InfoVis is a single track conference, we group papers within each year by session, so the ordering of sessions and papers are consistent with the actual conference program. Participants could browse through the proceedings or search for specific papers by title, author, or abstract.

The most scarce resource in conducting the survey was acquiring available expert time. To maximize the value of their responses, we chose exemplary words and documents as the means to express a concept. Labels are widely used in cognitive psychology (Rosch et al., 1976) for identifying topics. Based on pilot studies, the two chosen properties — freeform typing of a list of terms, and drag-and-drop specification of papers — minimize input complexity and allow experts to focus on the construction of topics. We omitted other descriptive attributes, such as summary sentences, which took pilot participants much longer to enter. We displayed twenty default boxes to provide reasonably exhaustive coverage of the domain while bounding the length of the survey. In a preliminary study, two of the authors exhaustively annotated every document in the corpus with multiple tags. The overlap between the two sets of annotations indicated that the domain was covered by approximately twenty shared topics.

2.2. Survey Data

We received a total of 202 topical responses (maximum of 22 and minimum of 18 per subject). The participants specified an average of 5.71 terms (max 19, min 1, median 8) and 5.15 documents (max 25, min 1, median 7) per topic. Subjects provided 171 distinct topic labels and 769 distinct terms. Together, the experts cited a total of 342 distinct documents (77% of all papers published at IEEE Information Visualization Conference) which we consider to be a reasonable coverage of the field.

We analyzed timing information for seven participants who had active internet connections for the full duration of their survey. The survey webpage automatically saved responses every minute, allowing us to track changes at that granularity. On average, the experts spent 91.7 minutes (max 162, min 42) edit-

ing their responses within a maximum of five sessions. The amount of editing time suggests that the survey taxed the experts attention and available contiguous time.

3. Mathematical Derivation: Convolution Operator

Given a Bernoulli process $\{x^i\}_{i=1,\text{definitive}}^{\infty}$ where x^i is the probability of observe a positive outcome for the i-th event, let $P^k_{\text{definitive}}(n)$ represent the probability that we observe exactly n positive outcomes among the following k events $\{x^i\}_{i=1,\text{definitive}}^{k}$.

Similarly, given a Bernoulli process $\{x^i\}_{i=1,\text{noise}}^{\infty}$ where x^i is the probability of observe a positive outcome for the i-th event, let $P^k_{\text{noise}}(n)$ be the probability that we observe exactly n positive outcomes among the following k events $\{x^i\}_{i=1,\text{noise}}^{k}$.

Suppose we construct a new series of Bernoulli process $\{x^i\}_{i=1,\text{combined}}^{m}$ consisted of m events, by randomly drawing from the two processes $\{x^i\}_{\text{definitive}}$ and $\{x^i\}_{\text{noise}}$. Suppose we draw k events from $\{x^i\}_{\text{definitive}}$ and $m-k$ events from $\{x^i\}_{\text{noise}}$.

Let $P_{\text{combined}}(n)$ be the probability that we observe exactly n positive outcomes among its m events. I claim that: $P_{\text{combined}} = P^k_{\text{definitive}} * P^{m-k}_{\text{noise}}$

3.1. Sampling from Two Bernoulli Processes

Since events in a Bernoulli process are considered independent, we can re-arrange the order of events without affecting the expected number of positive outcomes.

3.2. Sampling from Definitive vs. Noise Charts

When computing the expected number of positive outcomes for P_{combined}, the combined definitive and noise charts, we re-arrange the series $\{x^i\}_{\text{combined}}$ so that the k definitive events occur first and the $m-k$ noise events later.

3.3. Convolution

Let $\{x^i\}$ be a Bernoulli event where x^i is the probability of observing a positive outcome for event i. We construct a 2-vector $X^i = [1-x^i, x^i]^T$.

Let P^k be the multinomial distribution representing the observed cumulative outcome of the first k events where $P^k(n)$ is the probability that we observed exactly n positive outcomes for the first k events. We represent P^k as an $(k+1)$-vector with entries $[P^k(0), P^k(1), \cdots, P^k(k)]^T$.

We prove by induction, that $P^{k+1} = P^k * X^{k+1}$.

As the base case:

$$P^0 = [1]^T$$
$$P^1 = X^1 = 1 * X^1 = P^0 * X^1$$

For the inductive step:

$$\begin{aligned} P^{k+1}_i &= P^k_{i-1} \cdot x^{k+1} + P^k_i \cdot (1 - x^{k+1}) \\ &= P^k_{i-1} \cdot X^{k+1}_1 + P^k_i \cdot X^{k+1}_0 \\ &= \sum_{t=0}^{1} P^k_{i-t} \cdot X^{k+1}_t \\ P^{k+1} &= P^k * X^{k+1} \end{aligned}$$

3.4. Associativity

Let $P^{j,k}$ represent the observed cumulative outcome for events j to k (inclusive). Since convolution is associative:

$$\begin{aligned} P^{0,m} &= P^{0,k} * X^{k+1} * X^{k+2} * \cdots * X^n \\ &= P^{0,k} * P^{k+1,m} \end{aligned}$$

It follows that the expected topical-concept matches for the combined chart is:

$$P_{\text{combined}} = P_{\text{definitive}} * P_{\text{noise}}$$

4. Noise Estimation

4.1. Setting k

By construction, the distributions P and P_{noise} have the same mean. We arbitrarily choose k so that $P_{\text{definitive}}$ has the same mean as P.

For non-integer values of k, $P_{\text{definitive}}$ is zero everywhere except for two values, $P_{\text{definitive}}(\lfloor k \rfloor) = \lceil k \rceil - k$ and $P_{\text{definitive}}(\lceil k \rceil) = k - \lfloor k \rfloor$.

4.2. Solving for γ

Discrete convolution can be convert to matrix multiplication. We convert the "convolute by P_{noise}" operation into a Toeplitz matrix $A = A_{\text{noise}}$. Let $P' = P^{k(1-\gamma)}_{\text{definitive}}$.

$$\begin{aligned} &\operatorname*{argmin}_{\gamma} \mathrm{KL}(P' * P^{\gamma}_{\text{noise}} || P) \\ &\operatorname*{argmin}_{\gamma} \mathrm{KL}(AP' || P) \\ &\operatorname*{argmin}_{\gamma} P'^T A^T \log(AP') - P'^T A^T \log(P) \end{aligned}$$

We apply gradient descent to determine the optimal value γ that minimizes the (convex) objective function.

5. Denoise Computation

5.1. Solving for P_{denoised}

Let $P'' = P_{\text{denoised}}$.

$$\begin{aligned}
&\underset{P''}{\operatorname{argmin}}\, \mathrm{KL}(P'' * P_{\text{noise}} || P) \\
&\underset{P''}{\operatorname{argmin}}\, \mathrm{KL}(AP'' || P) \\
&\underset{P''}{\operatorname{argmin}}\, P''^T A^T \log(AP'') - P''^T A^T \log(P)
\end{aligned}$$

subject to

$$\begin{aligned}
&\sum_i P''_i = 1 \\
&0 \leq P''_i \leq 1 \quad \text{for all } i
\end{aligned}$$

The above is an optimization involving both equality and inequality constraints. We apply sequential quadratic programming to solve to P'' using three mathematical components: barrier method to remove inequality constraints, first-order trust region to solve for equality-constrained minimizations, and heuristics to obtain a good initial solution.

5.2. Outer Iteration: Barrier Method

We apply barrier method to remove the inequality constraints, in order to reduce complexity and speed up computation. We modify the objective function as the following.

$$P''^T A^T \log(AP'') - P''^T A^T \log(P) + e^{-\alpha P''} + e^{\alpha(1+P'')}$$

We perform 50 iterations and gradually increase α from 500 to 50000.

5.3. Inner Iteration: Trust Region

Within each iteration of the barrier method, we applied first-order trust region solve for an optimal solution P''.

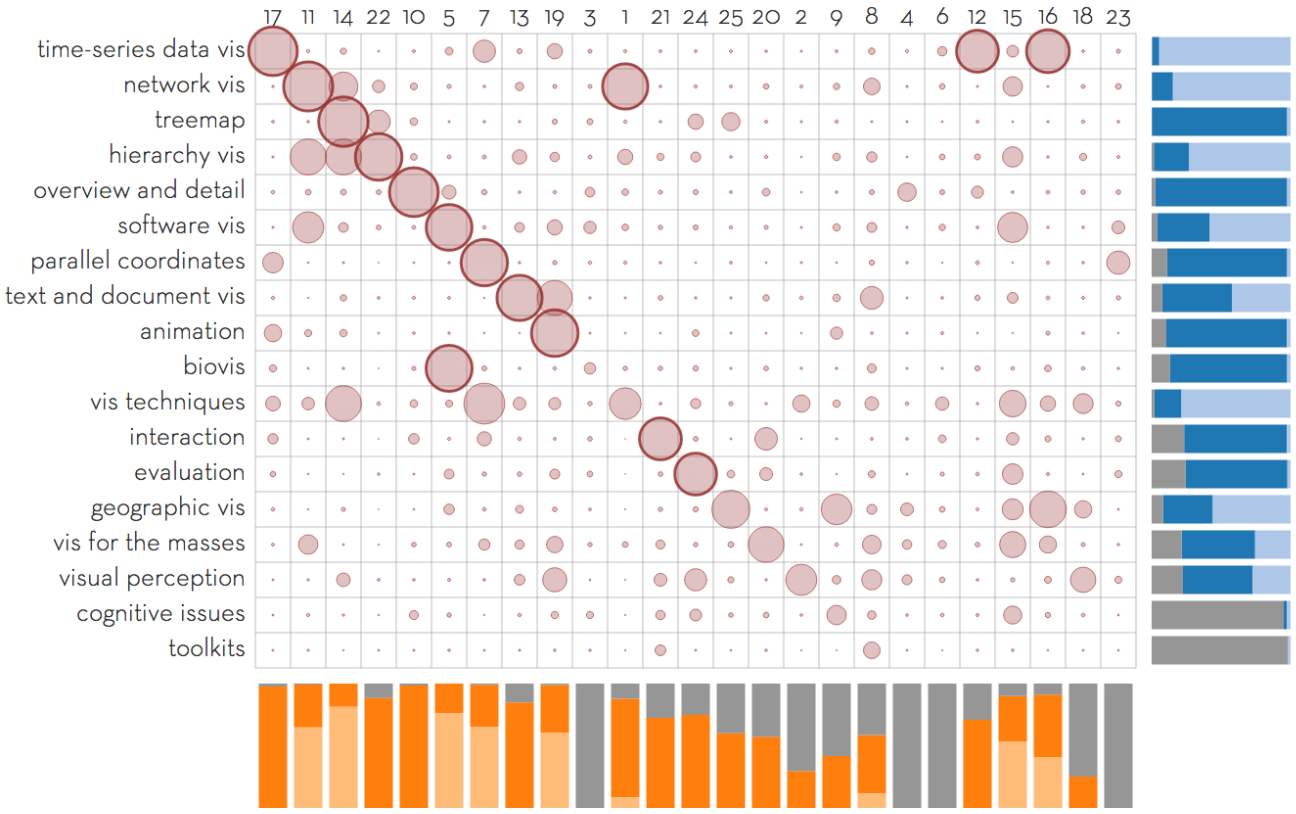
5.4. Initial Solution

To ensure better convergence, we solve the linear system of equations $AP'' = P$, to obtain an initial solution $P''^{(0)}$. We clamp the values of $P''^{(0)}$ to within $[0, 1]$ and L^1 normalize the vector to ensure it's a valid probability distribution. We use the resulting vector as the initial solution for the aforementioned barrier method/trust region solver.

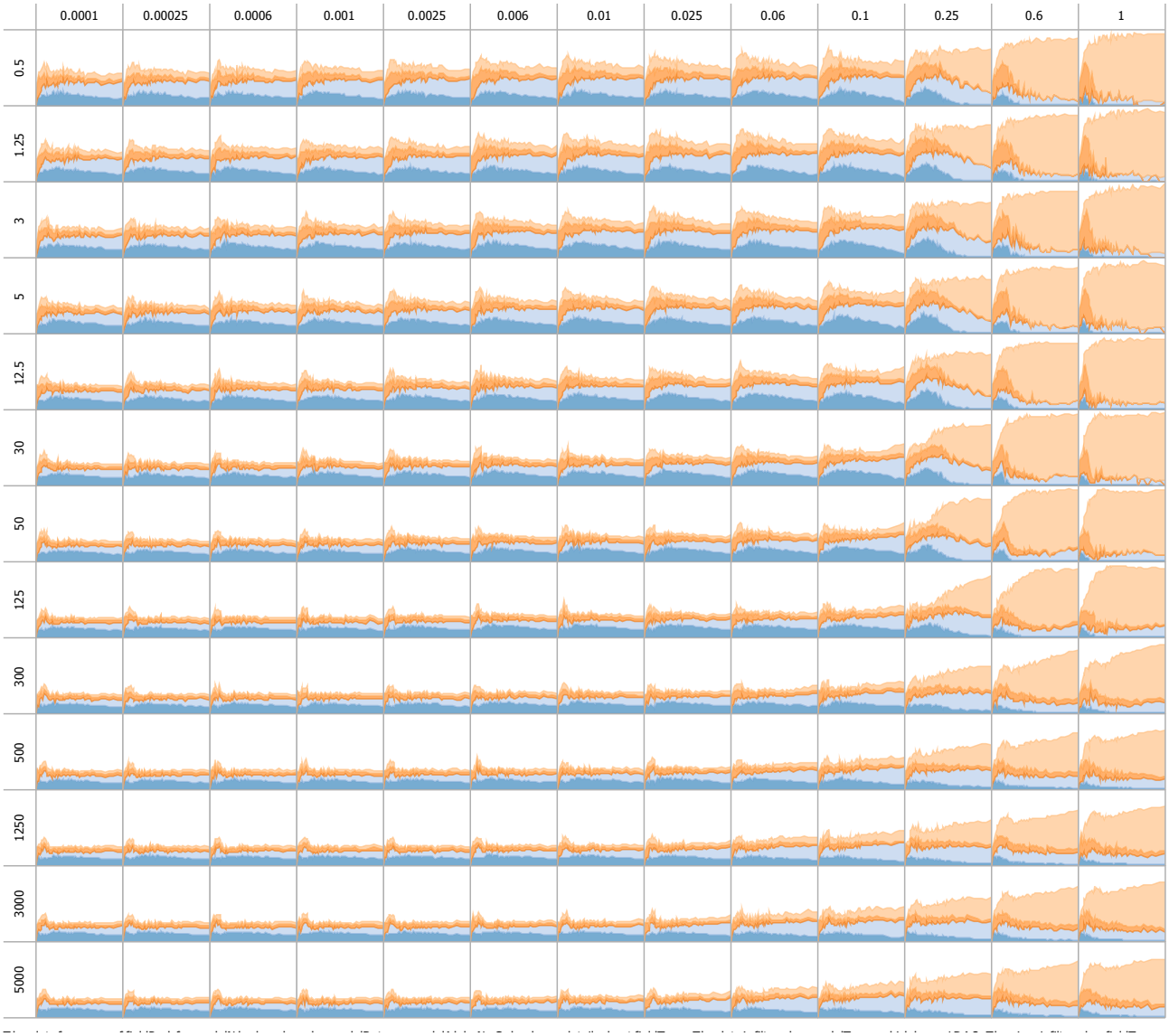
References

Griffiths, Thomas L. and Steyvers, Mark. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(1):5228–5235, 2004.

Rosch, Eleanor, Mervis, Carolyn B, Gray, Wayne D, Johnson, David M, and Boyes-Braem, Penny. Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439, 1976.



Supplementary Figure 2: *Correspondence chart between latent topics and reference concepts.* The set of 25 latent topics are generated by a LDA model ($N = 25, \alpha = 0.01, \beta = 0.01$) and displayed along the columns. The set of 18 reference concepts are given by one of the InfoVis experts and displayed along the rows. Area of circles represents the matching likelihood between topic-concept pairs; likelihoods exceeding random chance are marked with a bold border. Bars on the right show the probability that a concept is *missing* (grey), *resolved* (blue), or *repeated* (light blue). Bars on the bottom show the probability that a topic is *junk* (grey), *resolved* (orange), or *fused* (light orange). This visual analysis tool is available online at: `http://vis.stanford.edu/topic-diagnostics`



Supplementary Figure 3: *Exhaustive grid search.* Topical alignment for LDA models over a grid of parameter/hyperparameter settings: $N \in [1, 80]$ (horizontal axis across subgraphs), 13 values of $\alpha \in [0.5/N, 5000/N]$ (vertical axis), and 13 values of $\beta \in [0.0001, 1]$ (horizontal axis). We observe a qualitative shift in topical composition around β=0.25. For $\beta > 0.25$, the models generate fused topics that uncover but do not fully resolve a majority of the reference concepts as N increases. For $\beta < 0.25$, the proportion of resolved and fused topics remain stable regardless of N. Overall, decreasing β or increasing α leads to a decrease in coverage.