TopicPanorama: a Full Picture of Relevant Topics

Shixia Liu, *Senior Member, IEEE*, Xiting Wang, Jianfei Chen, Jun Zhu, and Baining Guo,*Fellow, IEEE*

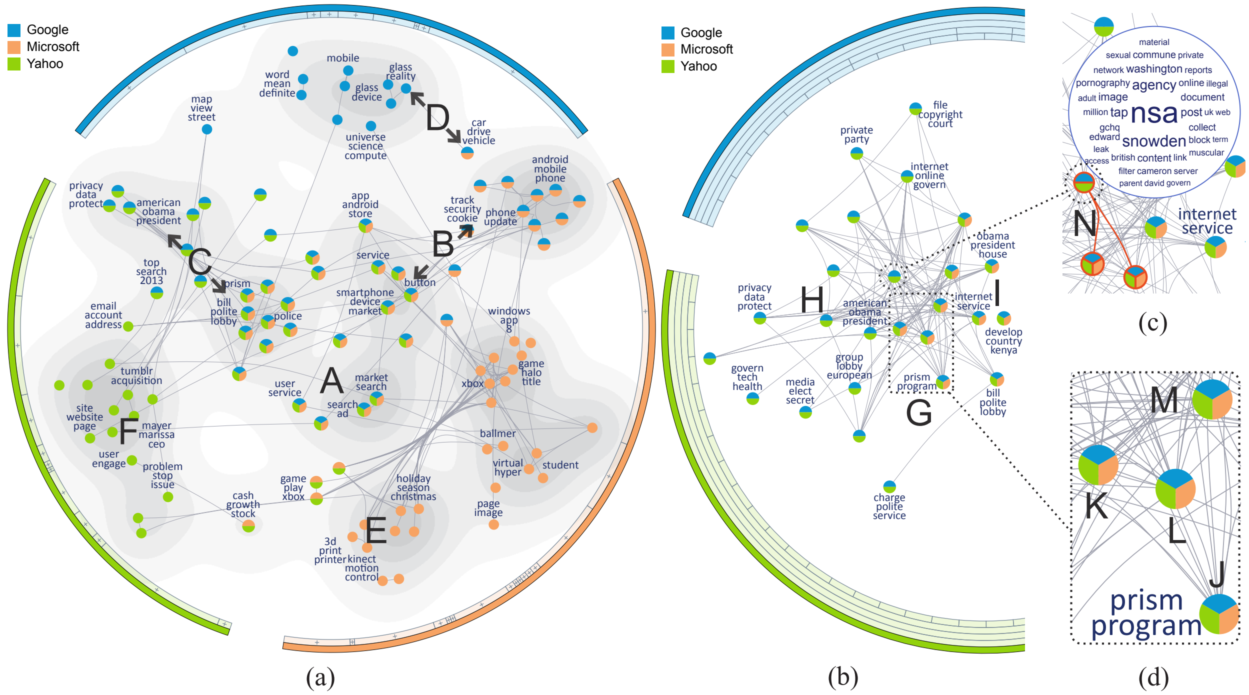


Figure 1. TopicPanorama visualization: (a) a full picture of topics related to Google, Microsoft, and Yahoo; (b) government related topics; (c) NSA Prism spying scandal shared by Google and Yahoo; (d) NSA Prism spying scandal shared by the three companies. Notations A-I represent different groups of topics and J-N represent different topics.

Abstract— We present a visual analytics approach to developing a full picture of relevant topics discussed in multiple sources such as news, blogs, or micro-blogs. The full picture consists of a number of common topics among multiple sources as well as distinctive topics. The key idea behind our approach is to jointly match the topics extracted from each source together in order to interactively and effectively analyze common and distinctive topics. We start by modeling each textual corpus as a topic graph. These graphs are then matched together with a consistent graph matching method. Next, we develop an LOD-based visualization for better understanding and analysis of the matched graph. The major feature of this visualization is that it combines a radially stacked tree visualization with a density-based graph visualization to facilitate the examination of the matched topic graph from multiple perspectives. To compensate for the deficiency of the graph matching algorithm and meet different users' needs, we allow users to interactively modify the graph matching result. We have applied our approach to various data including news, tweets, and blog data. Qualitative evaluation and a real-world case study with domain experts demonstrate the promise of our approach, especially in support of analyzing a topic-graph-based full picture at different levels of detail.

Index Terms—Topic graph, graph matching, graph visualization, user interactions, level-of-detail.

1 Introduction

A series of relevant topics, such as the US presidential election or a competitor/partner analysis of several related companies, is often heavily discussed in multiple sources such as news, blogs, or micro-blogs. These sources share a number of common topics while also having their own distinctive topics, which together form a full picture of relevant topics. A number of recent studies suggest that a better understanding of the full picture provides new insights for decision-making [45,61]. However, users often take great pains to develop a comprehensive understanding of the whole story. They have to repeatedly switch back and forth from one source to another in order to completely understand the full picture of given topics or events. For example, when responding to the NSA Prism spying scandal, a public relations manager needs to detect related topics in news articles about his company and examine their relationships with each other. Meanwhile, he also pays close attention to the reaction of other related companies, searching for corresponding topics in the news corpora of other companies and comparing them with each other. In performing this analysis, he often goes back and forth across multiple sources to compare topics of interest. To support such an analysis process, it is important to be able to gather separate pieces of information about these topics scattered in different sources and reconstruct the full picture.

- *Shixia Liu is with Microsoft Research Asia. E-mail: shliu@microsoft.com.*
- *Xiting Wang is with Tsinghua University and Microsoft Research Asia. E-mail: v-xitwan@microsoft.com.*
- *Jianfei Chen is with Dept. of Comp. Sci. & Tech., TNList Lab, State Key Lab of Intell. Tech. & Sys., Tsingua University. E-mail: chenjf10@mails.tsinghua.edu.cn.*
- *Jun Zhu is with Dept. of Comp. Sci. & Tech., TNList Lab, State Key Lab of Intell. Tech. & Sys., Tsinghua University. E-mail: dcszj@mail.tsinghua.edu.cn.*
- *Baining Guo is with Microsoft Research Asia. E-mail: bainguo@microsoft.com.*

Topic graphs have become a widely used approach for a more coherent understanding of a large collection of documents and easily finding information of interest [4, 11, 45]. These graphs are very important for illustrating how topics are connected to each other and providing an efficient but comprehensive understanding of topics of interest through correlation. As a result, a straightforward way of developing a full picture is to merge all data collected from different sources and then utilize a topic graph construction method, such as the correlated topic model (CTM) [4, 11], to build a topic graph on the merged data. However, there are two drawbacks to this approach. First, different text corpora contain texts with different lengths and language usages. For example, news articles are long and well formed, while tweets are short and noisy. This makes it difficult to use a unified topic graph generation method to build a single topic graph that fits each corpus well. Second, even when document lengths and language usages are similar, different corpora may have their own unique topics. Direct use of the topic graph construction method (with the same parameters) on all data may fail to model the diversity across different corpora because the model uses a common set of topics to model all data [45, 61]. We report the deficiency of using one topic graph construction method in Sec. 8.1.

To solve these issues, we have developed an interactive, visual analytics tool, called TopicPanorama. This tool aims to consistently integrate multiple topic graphs together to support iterative, progressive topic graph synthesis and analysis. We develop a multiple graph matching algorithm to find a consistent mapping among multiple topic graphs. Our algorithm is based on one of the most widely used pairwise graph matching metrics, graph edit distance [21, 43]. The major feature of the proposed graph matching algorithm is that it jointly optimizes related pairwise matches instead of performing a sequence of pairwise matches with a reference graph, which may introduce inconsistency. For example, a pairwise matching sequence may match v_2 to v_4 ($G_1 \mapsto G_2$) and v_4 to v_7 ($G_2 \mapsto G_3$) as illustrated in Fig. 4, but this conflicts with the direct matching result where v_2 matches v_9 ($G_1 \mapsto G_3$). Through the joint optimization, such inconsistencies are avoid. After graph matching, we employ a constrained multi-branch hierarchical clustering method to hierarchically organize each topic graph. The hierarchy provides an effective way to navigate large topic graphs. We then develop an LOD-based visualization for understanding the matched graph, which combines a radially stacked tree visualization with a density-based graph visualization. With this combination, TopicPanorama enables users to examine both the overarching concepts and fine details in each corpus. For example, it allows users to "zoom in" and "zoom out" to find specific or broad topics. Finally, we design a set of rich interactions to support the editing of the graph matching result and help analyze the matched graph. Since the graph matching algorithm is not perfect and different users may have different information needs, we allow users to interactively modify the results. TopicPanorama then updates the matching results accordingly. To this end, an incremental graph matching algorithm based on the Hungarian algorithm [36] is developed to integrate user feedback into the graph matching algorithm in real time.

The major contributions of this work are:

- **A visual analytics system** that helps users understand and analyze the full picture of relevant topics from different corpora.
- **A consistent multiple graph matching algorithm** that enables a joint optimization among topic graphs and supports real-time modifications of the matching result and a variety of interactions.
- **An LOD-based visualization** that allows users to understand and interact with the matched graph at different levels of granularity.

2 Related Work

2.1 Topic Visualization

Topic visualization, which aims to facilitate the understanding and analysis of text corpora based on topics, has received considerable attention in recent years [31, 48]. Generally, it can be classified into two categories: dynamic topic visualization and static topic visualization. Most of exiting dynamic topic visualizations focus on analyzing evolving topics based on a river metaphor. For example, Havre et al. [24] made an initial effort to employ a river metaphor to convey evolving topics over time. To help analysts better explore and analyze a large document collection, TIARA [32] tightly integrates the stacked graph visualization with the LDA model [5] to illustrate topic evolution patterns over time. Inspired by the same metaphor, visual Backchannel [18] was developed to visualize keyword-based topics that are extracted from tweets. ParallelTopics [19] employs ThemeRiver to illustrate topic evolution over time and parallel coordinate plots to convey the probabilistic distribution of a document on different topics. TextFlow [15] and RoseRiver [16] leverage Sankey diagrams to visually convey topic merging and splitting relationships over time. A visual analysis system was designed by Xu et al. [58] to allow analysts to interactively explore and understand the dynamic competition relationships among topics and the influence of opinion leaders. Recently, Sun et al. [49] extended this work to study both the cooperation and competition relationships among topics. The aforementioned approaches focus on the visual exploration of evolving topics from a single source. Different from these approaches, our work aims to provide a full picture of relevant topics from multiple sources.

Static topic visualizations leverage word lists or word clouds to visualize topic models. For example, Chaney and Blei [10] employs word lists to illustrate the hidden structure discovered by a topic model. This visualization is useful for users to understand the major topics in a document collection and the topic distribution in a document. However, it may fail to provide the correlations between topics and a full picture of many relevant topics from multiple corpora.

HierarchicalTopics [20] hierarchically organizes the extracted topics by the BRT model [6, 33] and thus can represent a large number of topics without being cluttered. However, HierarchicalTopics provides an overview of the topics extracted from one text corpora. While our method provides a full picture of relevant topics from multiple corpora and allows users to examine common topics among corpora as well as distinctive topics of each corpus. Technically, each corpus in our method is represented by a topic graph while the topics in HierarchicalTopics are organized by a tree.

Another work related to ours is FacetAtlas [9]. It also adopts the density-based graph visualization to represent the multifaceted relationships of documents within or cross the document clusters. However, it may fail to easily distinguish the common topics and distinctive topics across multiple corpora if we directly employ it to visualize the matched topic graph.

2.2 Graph Matching

A large number of graph matching methods have been proposed [14, 44]. Most of them focus on finding correspondence between two graphs [43, 44]. In this section, we only review error-tolerant graph matching methods since they can flexibly accommodate the differences between graphs by relaxing matching constraints. Such relaxation is very useful for topic graph matching, which often matches related graphs rather than exactly the same ones. Existing error-tolerant pairwise graph matching methods can be classified into the following categories: graph edit distance [21, 43], artificial neural network [47], relaxation labeling [54], spectral method [55], and graph kernel [40]. Among them, the most commonly used method is based on the edit distance of graphs [21, 39, 43, 44]. The basic idea of this method is to measure the structural difference of graphs by the number of edit operations needed to transform one graph into another.

Although these methods work well for matching two graphs, they are not suitable for matching three or more graphs. Directly using pairwise matching methods to match multiple graphs may introduce inconsistency [59]. Simply removing the inconsistent results may lead to suboptimal results, as detailed in Sec. 4.

To tackle this issue, there have been some efforts to match multiple graphs. Williams et al. [53] presented a proof-of-concept for multiple graph matching. They adopted a Bayesian framework to construct an inference matrix and used it to measure the mutual consistency of multiple graph matching. The framework looks promising, but no solver is provided, which limits the scope of deployment for this method. To compute a representative of a set of graphs, a common labeling algorithm [41, 42] has been developed. The algorithm learns common labels through a consistent multiple isomorphism. It can find a consistent common labeling among multiple graphs. However, it assumes that each graph has the same number of nodes.

Yan et al. [59] provided a multiple graph matching algorithm based on the pairwise matching solver and constrained integer quadratic programming (IQP). However, IQP is known to be computationally expensive, which makes this algorithm not applicable for real-time interactions. Furthermore, it may fail to infer mapping relationships among non-common parts of graphs (Sec. 4). Compared with [59], our method addresses the bottleneck of computation and missing mappings. We formulate multiple graph matching as a unified optimization approach based on graph edit distance and the Hungarian algorithm [36]. The inconsistency is resolved by seamlessly integrating direct pairwise mapping into a mapping formulation in a manner that direct maps can be distinguished from indirect maps. By leveraging an incremental Hungarian algorithm [28], our method allows users to interactively modify the matching results. We also develop an LOD-based visualization to better understand and analyze the matched graph from multiple perspectives.

2.3 Visual Graph Comparison

Visual graph comparison aims to analyze the similarities and differences between graphs [22, 51]. A number of graph comparison methods have been proposed, which can be categorized into three general approaches: animated views, juxtaposition, and superposition.

The animated views approach uses animated node-link diagrams to convey the changes of a dynamic graph [2, 7, 17, 29]. Basically, the approach generates a sequence of graphs for each time point. Successive layouts of similar graphs should have minimal changes (stability). Furthermore, each of such layouts should still effectively convey the properties of the underlying graph (readability). It also animates the layout from one step to the next to help the viewer easily follow changes.

Juxtaposition presents two graphs side-by-side temporally or spatially [3, 38]. Typical examples include VisLink [13], which displays each visualization in its own 2D plane and reveals connecting relationships by drawing links between them. Bremm [8] developed a visualization toolkit to compare multiple trees globally and locally. To make this practical, they presented only a few trees at a time for side-by-side comparison. Since each graph will be fully displayed, Juxtaposition may not scale well to large graph comparison.

Superposition combines multiple graphs into a bigger one and then places this graph with the same layout. Several methods have been proposed to overlay several graphs together for a variety of comparison related tasks. Alper et al. [2] overlaid two matrices or two node-link diagrams together to support pairwise weighted graph analysis. Vehlow et al. [50] developed a visualization technique to help users compare and analyze overlapping communities in networks. The LOD technique was adopted to support the investigation of fuzzy communities from a fully aggregated graph to the original graph.

The above methods assume that there is an exact matching between the corresponding nodes of different graphs. This may limit their applications since most graphs in real-world applications do not have such a correspondence. To bridge the gap, some recent efforts have begun to introduce graph matching techniques into visual graph comparison. For example, Sambasivan et al. [46] introduced a pairwise graph matching technique into the comparison of request-flow. They used heuristics to extract approximate matching between graphs. Hascoët et al. [23] developed an interactive graph matching tool that combines node-link diagrams with graph matching techniques. A heuristic rule based on the layout positions of nodes was used to approximately match nodes from different graphs. Although the matching method is simple and easy to implement, it may introduce more errors/uncertainties since the node position is not a reliable metric to match nodes. The adopted layout method does not distinguish between common and distinctive topics perceptually. Furthermore, the direct use of the force-directed layout cannot scale to large graphs. Compared with this method, TopicPanorama consistently integrates multiple topic graphs together to form a full picture of relevant topics, based on their content and relationships with each other. Specifically, we have developed a consistent multiple graph matching algorithm and tightly integrated it with an LOD-based visualization. Our LOD visualization combines a radially stacked tree visualization with a density-based graph visualization, which enables users to easily see the matching result, including the matched graph as well as individual ones.

3 TopicPanorama

3.1 Task Analysis

We developed TopicPanorama through multiple participatory design sessions with a group of experts, including two public relations managers, two journalists, and two sociologists. All participants were self-identified as having analytical experience in forming a full picture of relevant topics from multiple sources. They usually formed the picture by manually analyzing all the available documents, which is very time-consuming and requires high expertise. The experts expect a toolkit that allows them to effectively conduct analysis on a much larger dataset and can greatly advance their understanding of a full picture of the relevant topics of interest. In the design sessions, we focused on probing the participants' analysis processes and needs iteratively. We identified the following high-level tasks by close collaboration with these experts and iteratively conducting the nested model for visualization design and validation [37].

T1 - understanding the full picture of relevant topics. All experts expressed the need to smoothly navigate a full picture when analyzing relevant topics that are discussed in multiple sources, from the high-level topics to the detailed documents. The information of relevant topics is often scattered across multiple media sources. The experts often have to examine two or three corpora and repeatedly switch back and forth from one source to another in order to see the full picture. They stated that they can benefit from a toolkit that can consistently integrate two or three sources in practice. This is consistent with the conclusion of previous experiments, namely that about four objects can be tracked in visual comparison [26, 60].

T2 - examining common topics and distinctive topics of each source. When analyzing a full picture, the experts often compare topics across sources, including the common and distinctive ones. To better understand the common and distinctive parts of different sources, the experts required the ability to examine the common topics across multiple corpora as well as the distinctive topics of each corpus in the same view.

T3 - examining the correlations between topics. All the experts wanted to understand the correlations between topics, especially the correlations between common topics and distinctive topics of each source, because such correlations help them find information of interest more quickly. For example, one sociologist commented, "When analyzing media framing of events, I need to understand how two discursive spaces (i.e., mass media and grass roots) interact with each other."

T4 - exploring the full picture at different levels of granularity. In real-world applications, a source may contain hundreds or even thousands of topics. Quickly getting an overview of these topics and then drilling down to the detailed content gradually is a very important step for the experts to perform various analysis tasks. For example, one public relations manager said, "In my daily work, I often process multiple sources that contain thousands of topics. It is very time-consuming and tedious to examine these topics one by one. As a result, I eagerly expect a toolkit that efficiently organizes a large number of topics in each source and extracts overarching high-level concepts to globally represent that source. Then I could select the topics of interest for further exploration."

3.2 System Overview

To help users to better perform the tasks described in Sec. 3.1, we have developed TopicPanorama. It contains the following features.

- Leveraging a topic graph to represent each source and hierarchically organizing the topic graph (**T3**, **T4**);
- Matching multiple topic graphs to form a full picture (**T1**);
- Placing the common parts near the area of each related source and the distinctive parts in the corresponding area of each source (**T2**);

Accordingly, TopicPanorama consists of three major modules: graph matching, hierarchy building, and an LOD-based graph visualization (Fig. 2). Given several topic graphs, the graph matching module generates consistent mappings among them. To handle large topic graphs effectively, the hierarchy building module generates a topic hierarchy based on the constraint-based tree clustering method proposed by Wang et al. [52]. The graph matching results and the topic hierarchies are then fed to the visualization module, which combines a radially stacked tree visualization with a density-based graph visualization to illustrate the graph matching results. Users can interact with the generated visualization for further analysis. For example, the user can modify one of the matching results, then TopicPanorama will incrementally update the matching results.

Fig. 3 depicts the user interface of TopicPanorama. It contains three different interaction areas: TopicPanorama visualization (Fig. 3(a)), control panel (Fig. 3(b)), and information panel (Fig. 3(c)). The visualization view provides an overview of the relevant topics across multiple corpora. It contains two parts: the stacked tree visualization to show the hierarchical structure of each topic graph and the density-based graph visualization to show the correlations between topics. The information panel will display the corresponding topic information of a selected topic, including the keyword description of each topic and the side-by-side topic keyword comparison to illustrate topic matching results. It also shows the corresponding documents of each selected topic to help better understand the topic content. The control panel consists of a set of controls that allow users to examine the uncertain matched topics and edit the matching results.

4 Consistent Graph Matching Algorithm

In this section, we study the problem of finding correspondence among multiple topic graphs.

4.1 Model

Graph edit distance is a widely used metric in graph matching algorithms to match two graphs [21, 43]. It measures the structural difference of graphs by the number of edit operations (e.g., *node insertion*, *deletion*, and *substitution*) needed to transform one graph into another.

Given two graphs $G_1 = (\mathcal{V}_1, \mathcal{E}_1)$ and $G_2 = (\mathcal{V}_2, \mathcal{E}_2)$, where $\mathcal{V}_1$, $\mathcal{V}_2$ are the node sets and $\mathcal{E}_1$, $\mathcal{E}_2$ are the edge sets, we denote the matching between them as $f_{G_1G_2}$. The graph edit distance between G_1 and G_2 is defined as the minimal cost of all edit paths between them:

$$d(G_1, G_2) = \min\ c(f_{G_1G_2}), \qquad c(f_{G_1G_2}) = \sum_{o_i} c(o_i), \tag{1}$$

where $c(f_{G_1G_2})$ is the edit cost that maps G_1 to G_2 and $c(o_i)$ denotes the cost function of the edit operation o_i.

Given N graphs, a natural extension of the bipartite matching method for multi-graph matching is to summarize the graph edit distance of each pairwise matching (Baseline 1), that is,

$$d(G_1, G_2, ..., G_N) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} d(G_i, G_j). \tag{2}$$

However, this formulation may introduce inconsistency into the mappings. Fig. 4 shows an example. The three topic graphs are generated by applying CTM [4, 11] on three news corpora related to Yahoo (G_1), Microsoft (G_2) , and Google (G_3). CTM is a very effective method to learn topics as well as their correlation structure by employing a logistic-normal prior in a hierarchical topic model [4]. In this figure, nodes with the same color are about the same topic, which is labeled by the expert. For example, blue nodes are about an Indian court summoning 21 websites (e.g., Google, Facebook) for objectionable content and the purple nodes are about the US presidential election in 2012. $f_{G_1G_2}$ maps v_2 to v_4. $f_{G_2G_3}$ maps the two purple nodes together ($v_4 \mapsto v_7$). Here $v_i \mapsto v_j$ indicates that v_i is mapped to v_j. From these two mappings, we note that node v_2 maps to node v_7, which conflicts with the direct mapping result of $f_{G_1G_3}$ ($v_2 \mapsto v_9$). Similar inconsistency is observed in the mappings among the nodes v_3, v_5, v_8, and v_{10}.

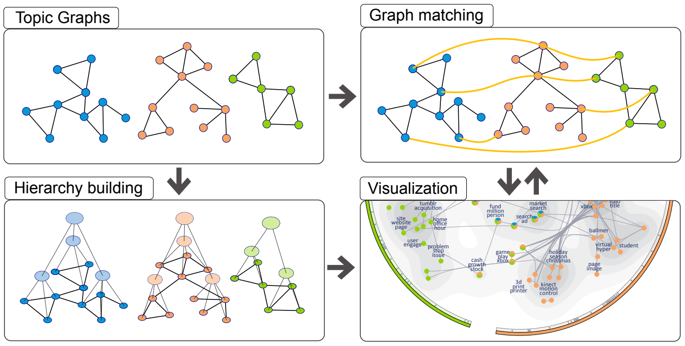
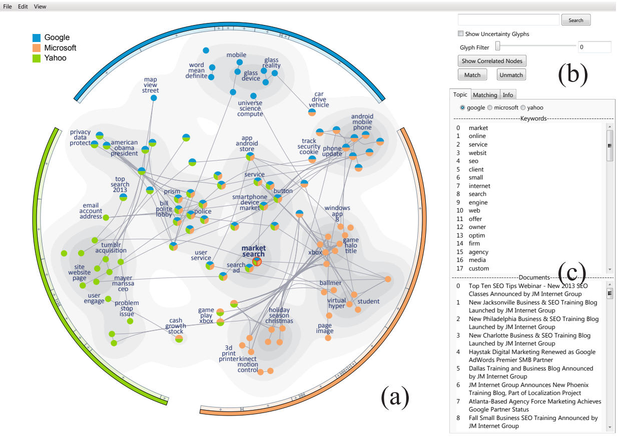


Figure 2. TopicPanorama overview.



□ Show Uncertainty Glyphs
Glyph Filter 0

Topic | Matching | Info

● google ○ microsoft ○ yahoo

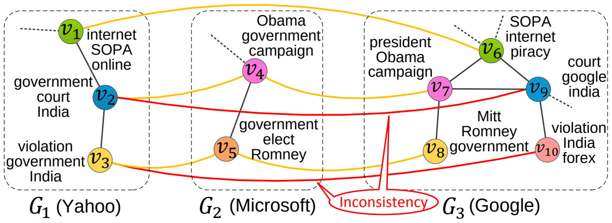
---Keywords---

0 market
1 online
2 service
3 websit
4 seo
5 client
6 small
7 internet
8 search
9 engine
10 web
11 offer
12 owner
13 optim
14 firm
15 agency
16 media
17 custom

---Documents---

0 Top Ten SEO Tips Webinar - New 2013 SEO Classes Announced by JM Internet Group
1 New Jacksonville Business & SEO Training Blog Launched by JM Internet Group
2 New Philadelphia Business & SEO Training Blog Launched by JM Internet Group
3 New Charlotte Business & SEO Training Blog Launched by JM Internet Group
4 Haystak Digital Marketing Renewed as Google AdWords Premier SMB Partner
5 Dallas Training and Business Blog Announced by JM Internet Group
6 JM Internet Group Announces New Phoenix Training Blog, Part of Localization Project
7 Atlanta-Based Agency Force Marketing Achieves Google Partner Status
8 Fall Small Business SEO Training Announced by JM Internet Group

Figure 3. User interface: (a) TopicPanorama visualization; (b) control panel; (c) information panel.



v_1 internet SOPA online
government court India v_2
violation government India v_3
Obama government campaign v_4
government elect Romney v_5
SOPA internet piracy v_6
president Obama campaign v_7
Mitt Romney government v_8
court google india v_9
violation India forex v_{10}
G_1 (Yahoo)
G_2 (Microsoft)
Inconsistency
G_3 (Google)

Figure 4. Inconsistency caused by directly applying pairwise matching to multi-graph matching. The node mappings between G_1 and G_3 derived from the matching results of $f_{G_1G_2}$ and $f_{G_2G_3}$, $v_2 \mapsto v_4$ and $v_4 \mapsto v_7$, are inconsistent with the direct matching result of $f_{G_1G_3}$, $v_2 \mapsto v_9$. Here $v_i \mapsto v_j$ indicates that v_i is mapped to v_j.

A simple strategy to resolve inconsistency is to remove the conflicted nodes. However, many inconsistencies may exist and it is difficult to find the optimal solution, especially when the number of conflicted nodes is large. Another option is to treat the pairwise matching results that can be derived from other matching results as constraints and add them to the related pairwise matching procedures to ensure consistency between different graph matching results (Baseline 2). By doing so, consistency between the common matched parts across all the graphs is obtained. However, it may fail to infer the mapping relationships among the non-common parts of the graphs. Fig. 5 shows two pairwise matching results for three graphs. Although the matching is consistent, we fail to infer the mapping relationships between nodes that do not have a corresponding node in G_2. For example, we do not know whether v_1 maps to v_6, v_9, or another node because v_1, v_6 and v_9 do not have a related node in G_2.

To solve this issue, we develop a consistent graph matching method that aims to minimize the cost of all pairwise graph matchings, with the constraint that all node mapping relationships are transitive. By ensuring such transitive relationships (consistency constraint), the proposed method derives globally consistent mappings across multiple graphs. Mathematically, the proposed graph matching method is formulated as

$$\begin{aligned} & d(G_1, G_2, ..., G_N) = \min\ c(f_{G_1G_2...G_N}), \quad c(f_{G_1G_2...G_N}) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} c(f_{G_iG_j}) \\ & s.t.\ v_l \mapsto v_m, v_m \mapsto v_n \Rightarrow v_l \mapsto v_n \\ & \quad \forall G_i, G_j, G_k \in \{G_1, G_2, ..., G_N\},\ \forall v_l \in \mathcal{V}_i, \forall v_m \in \mathcal{V}_j, \forall v_n \in \mathcal{V}_k, \end{aligned} \tag{3}$$

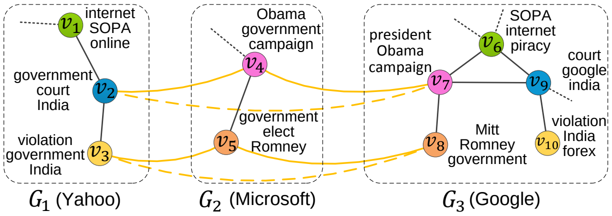


Figure 5. An example of simply adding constraints to pairwise matching. The matching result of $f_{G_1G_3}$ is regarded as the constraint of pairwise matching $f_{G_1G_2}$ and $f_{G_2G_3}$. There is no correspondence between v_1 and v_6 because they do not have a related node in G_2.

Next, we rewrite the cost function in Eq. (3) as

$$c(f_{G_1G_2\ldots G_N}) = \sum_{i=1}^{N-1}\sum_{j=i+1}^{N-1} c(f_{G_iG_j}) + \sum_{i=1}^{N-1} c(f_{G_iG_N}) = c(f_{G_1G_2\ldots G_{N-1}}) + \sum_{i=1}^{N-1} c(f_{G_iG_N}). \quad (4)$$

We then introduce the concept of *meta-graph* to further simplify the cost function. The meta-graph is constructed by merging the mapped nodes (or edges) as a meta-node (or meta-edge). The meta-graph is comprised of the consistently matched results of N graphs that contain both the common topics and distinctive topics of each topic graph. Fig. 6(a) shows an example of a meta-graph $M(G_1G_2)$ for the matching $f_{G_1G_2}$. When matching a meta-graph and a normal graph, we define the cost of each edit operation of a meta-node as the sum of the cost that maps each node in the meta-node to the normal node. Accordingly, the cost function of matching a meta-graph and a normal graph is

$$c(f_{M(G_1\ldots G_{N-1})G_N}) = \sum_{i=1}^{N-1} c(f_{G_iG_N}). \quad (5)$$

With this formulation, Eq. (4) is rewritten as

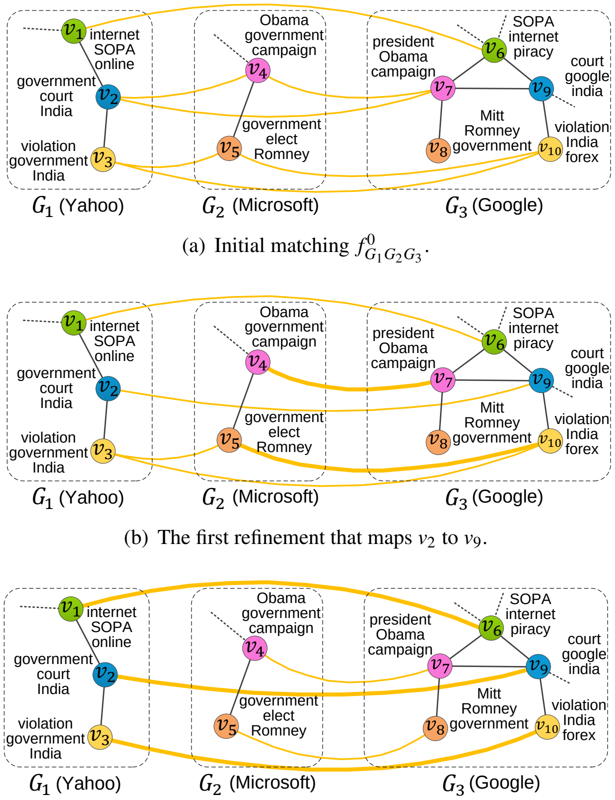
$$c(f_{G_1G_2\ldots G_N}) = c(f_{G_1G_2\ldots G_{N-1}}) + c(f_{M(G_1G_2\ldots G_{N-1})G_N}) \quad (6)$$

From the aforementioned formulation, we can derive the meta-graph of N matched graphs (i.e., the matching $f_{G_1G_2\ldots G_N}$) from the meta-graph of the $N-1$ matched graphs. Fig. 6(b) shows an example where we build the meta-graph $M(G_1G_2G_3)$ based on $M(G_1G_2)$.

4.2 Algorithm

Directly optimizing Eq. (6) is intractable; thus we employ an iterative greedy method to find an approximate solution. For $\forall k, 2 < k \leq N$, we first generate an initial consistent matching $f_{G_1G_2\ldots G_k}$ by directly mapping the optimal meta-graph $M(G_1G_2\ldots G_{k-1})$ of the previous step to the new graph G_k. Then in the refinement step, for each $1 \leq i < k$, we fix the mapping result of $f_{G_1\cdots G_{i-1}G_{i+1}\cdots G_k}$ and treat it as a meta-graph. Next, we map the meta-graph to G_i. If the cost of the new mapping is less than the old cost, we use the new mapping to replace the old one.

We use a simple example that contains three graphs to illustrate the basic idea of the algorithm. Fig. 7(a) shows the initial matching. Unlike the baseline method 2 (Fig. 5), our method can find the correspondence between v_1 and v_6. The initial matching is not optimal because G_3 is not considered when matching G_1 with G_2. For example, the blue node v_2 is incorrectly mapped to the purple node v_4. This incorrect mapping may cause more errors in the matching process that occurs later. For example, v_2 is incorrectly mapped to the purple node v_7 when matching $M(G_1G_2)$ with G_3. To solve this problem, we then iteratively refine the initial matching to get an optimal one. Fig. 7(b) shows the matching result of the first refinement. In this step, the matching between G_2 and G_3 is fixed as a meta-graph and the matching results between G_1 and G_2 as well as G_1 and G_3 are updated. The thick curves represent the fixed mappings and the thin curves represent the updated ones. Because the fixed mapping $v_4 \mapsto v_7$ is correct, $f^1_{G_1G_2G_3}$ correctly maps v_2 to v_9 by using this information. Fig. 7(c) shows the final matching result $f^2_{G_1G_2G_3}$ generated by the second refinement. Given that v_3 is mapped to v_{10}, $f^2_{G_1G_2G_3}$ correctly maps v_5 to v_8.



(a) Initial matching $f^0_{G_1G_2G_3}$.

(b) The first refinement that maps v_2 to v_9.

(c) The final result generated by the second refinement. v_5 is mapped to v_8.

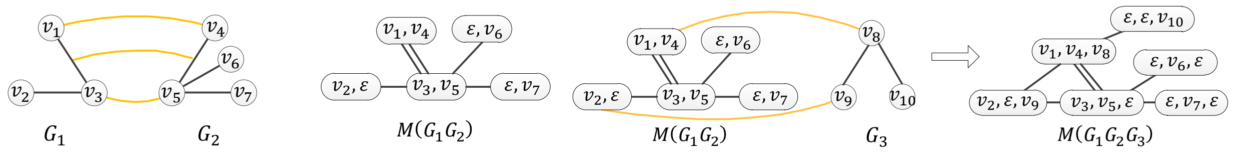
Figure 7. Illustration of our consistent graph matching algorithm.

4.3 Incremental Algorithm

Although the proposed graph matching method can successfully generate an optimal matching among multiple graphs, it may still be imperfect. Furthermore, different users may have different information needs. Thus one graph mining model cannot meet all possible requirements. To compensate for this, TopicPanorama allows users to interactively modify the graph matching result. Accordingly, an incremental graph matching algorithm is developed based on the incremental Hungarian algorithm [28], which easily integrates user feedback into the graph matching algorithm and updates the related mapping results in real time.

5 Panorama Visualization

The visualization design was guided not only by known perceptual principles (e.g., Gestalt laws), but also by the analysis needs of domain experts (Sec. 3.1) as well as their explicit feedback.



(a) $f_{G_1G_2}$ and its corresponding metagraph

(b) Derive the meta-graph $M(G_1G_2G_3)$ from $M(G_1G_2)$.

Figure 6. An example of meta-graph and its iterative matching operation. ε represents a null node.

5.1 Graph Matching as Density-Based Graph Visualization

Previous research has shown that a familiar visual representation lowers the cognitive load imposed on a user and benefits the learning process by employing the user's knowledge and experience [35]. Thus, the basic principle of our design is to employ a familiar visual metaphor when appropriate. We also employ a superposition comparison because this design is more efficient for comparing multiple graphs [2, 22]. Inspired by these two principles, we first considered a straightforward design that uses a single node-link diagram for two or three graphs. We adopt the node-link diagram instead of a matrix representation because the experts expressed the need to examine the correlations between topics (**T3**) and the node-link diagram is more intuitive than the matrix on showing the relationships between nodes [25]. This design was then presented to our target users for evaluation. Overall, they liked the overlaid design that provides a global overview first. Their major concerns were visual clutter caused by merging multiple graphs and scalability. One of them commented that, "This visualization helps me quickly get an overview of the topics that I am interested in. However, it is difficult for me to identify and analyze individual unique or common topics even though this visualization only contains 70 or 80 nodes."

To allow users to navigate and compare large topic graphs efficiently, we build hierarchies for topic graphs based on the Bayesian Rose Tree (BRT) model [33] (**T4**). We then develop a density-based graph visualization that combines a node-link diagram with a density map to display the nodes at the selected level of the topic hierarchies (Fig 1(a)). Specifically, we extract representative nodes for each of the topic nodes at the selected tree level and assign other non-representative nodes to their closest representative nodes. As shown in Fig. 1(a), the representative nodes are displayed as a node-link diagram and the other nodes as the density map. In this visualization, the common parts are placed near the area of each related corpus and the distinctive parts are placed in the area corresponding to the corpus (**T2**). The topic nodes of different corpora are encoded by different colors and the ones in common are represented by a pie chart (Fig. 1) with each of the slices corresponding to the matched corpus (**T2**). The node-link diagram is utilized to explain the relationships between representative nodes and the density is employed to illustrate global context (**T1**,**T3**). This design was well received by our target users. They all liked the hybrid visualization design in which both focus and context are well conveyed.

After engaging with the first prototype, the users identified some incorrect mapping results. They expressed the need to be prompted with an explicit request to examine such uncertain mapping results. This requirement is consistent with the conclusion of previous work that effectively conveying uncertainty in the matching results is very important to the data analysis process [34,56,57]. As shown in Fig. 9, we design a glyph to represent the uncertainty mappings with larger cost values. This glyph design is inspired by the iconic symbol called *filled bar and slider*, which is one of the intuitiveness winners for representing attribute uncertainty [34]. In this metaphor, we use the angle between the two sliders to encode the degree of uncertainty. A large angle indicates a high degree of uncertainty. Furthermore, we also allow users to interactively modify the mapping results according to their knowledge and needs. An incremental Hungarian algorithm [28] is employed to incrementally update the related mapping results locally based on user feedback.

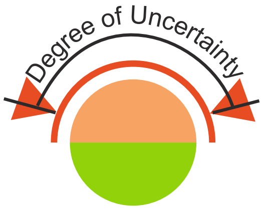


Figure 9. Uncertainty glyph.

5.2 Topic Hierarchy as Stacked Tree

To handle a large corpus with a large number of topics, we build hierarchies for topic graphs based on BRT [33] with each non-leaf node representing a topic cluster. The BRT model greedily estimates the tree structure with higher marginal likelihood. It can produce trees with arbitrary branching structure at each node. The detailed steps are described in Sec. 7. We employ radial, stacked tree visualizations (Fig 1(a)) to display topic hierarchies (**T4**). They are placed on the circumference of the radial layout, with the sector angle encoding the topic number of the corpus.

5.3 Coupling Graph Visualization with Stacked Tree

The two visualizations are integrated in a circular layout. The stacked trees form the boundary. The inner part of the radial layout is the density-based node-link diagram. These two visualizations are synchronized together to help users navigate large topic graphs from a global overview to local details (**T1**). For example, when a user selects a topic node from one stacked tree, its children are displayed and the other trees will update accordingly by displaying the children of the mapped topics. The density-based graph visualization also smoothly zooms into detailed topics.

6 Layout Algorithm

Given *N* corpora, the layout of the radial stacked tree is quite straightforward. We put the unique nodes and common nodes of the *N* corpora in the middle of the corresponding arc. Other common nodes that are mapped to fewer than *N* corpora are placed on a part of the arc that is close to the related tree nodes in other corpora. Next, we introduce the layout method of the density-based graph visualization. The basic principle of the layout is that the common parts are placed near the layout area of each corpus (corpus area) and the distinctive parts are placed in the related corpus area. For example, the common parts of all corpora are placed in the center of the layout area. The common parts of Corpora **A** and **B** are placed in-between the two related corpus areas (Fig. 8). In each part, the topic nodes under the same parents should be placed together (cluster-aware layout). To satisfy the aforementioned principle, we combine Voronoi tessellation with a force-directed graph layout [27].

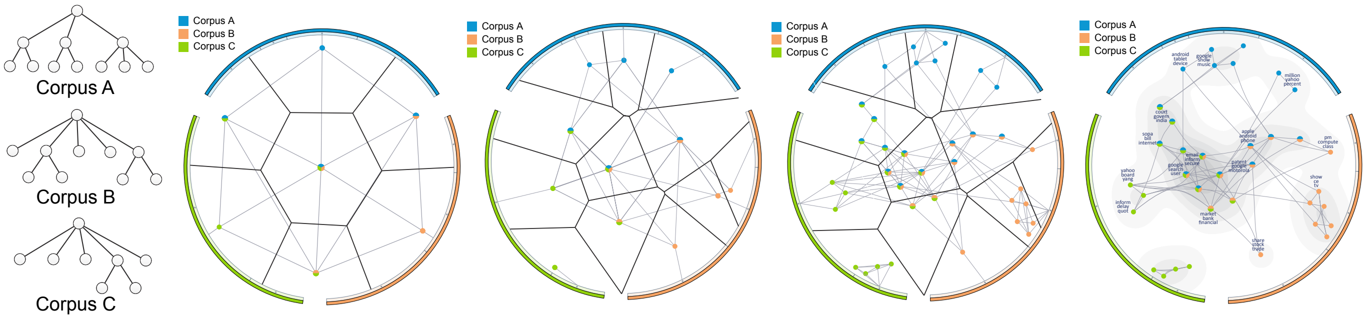
The first step involves deriving the layout centers of the common and distinctive parts in each corpus, respectively. The basic idea is to employ the force-directed graph layout method to compute the center position of each part. To this end, we build a graph according to the relationships between individual parts as well as the relationships between the staked trees and each part. As shown in Fig. 8(b), the common part has connections with each of the unique corpora it contains. The distinctive parts directly connect to the topic hierarchies (stacked trees) that they belong to. Next, the graph is laid out using the force-directed model, which provides the center position of each layout area. Based on center positions, a Voronoi tessellation is computed to allocate the layout area for each part. Within each layout area, we then place the topic nodes at the selected tree level. Based on the calculated node position, we compute another Voronoi tessellation (Fig. 8(c)). For each topic node at the selected tree level, we extract several representative leaf topics to represent the content of this node. We follow the topic ranking techniques, namely, coverage and variance as well as distinctiveness, proposed in TIARA [32], to select the representative leaf topics. For each selected representative topic, in addition to the connections in the topic graph, we also add a connection between each topic and the tree node it belongs to. With these added connections, the leaf topics are placed as close as possible to the tree node they belong to. Naturally, the leaf topics that belong to the same tree node are placed in the corresponding tessellation cell by a force-directed layout, which maintains the cluster structure among topics (Fig. 8(d)). In the third step, we assign each hidden leaf topic to the closest representative leaf topic and utilize kernel density estimation [30] to visually illustrate the global cluster context (Fig. 8(e)).

7 Implementation

In this section, we present several implementation details.

Construction of topic graph. In our implementation, we employ two methods to construct topic graphs: scalable correlated topic model (CTM) for long documents such as news articles and coupling scalable CTM with a postprocessing for short documents that includes linkages between documents (e.g., tweets that include co-hashtag and retweet relationships).

Scalable CTM. We adopt the very recent work of scalable CTM, which presents a scalable Gibbs sampling algorithm [11] and manages to learn



(a) (b) (c) (d) (e)

Figure 8. The basic idea of the layout algorithm: (a) topic hierarchies; (b) layout of the common and distinctive parts and compute the corresponding Voronoi tessellation; (c) layout of the cluster nodes of the selected tree level within each generated tessellation cell and compute a new Voronoi tessellation based on the new layout result; (d) Layout of representative nodes; (e) the final layout result.

	Dataset A			Dataset B			Dataset C		
	News	Blogs	BBS	Google	Microsoft	Yahoo	Baidu	Alibaba	Tencent
Separate	**2898.7**	**3792.6**	**2333.6**	**2604.9**	**2203.3**	**1822.5**	**2017.6**	**2022.3**	**2031.6**
Joint	3037.4	4058.4	2640.7	2872.2	2444.2	2202.7	2055.3	2183.4	2093.2

Table 1. Perplexity of topic models trained separately and jointly. The separately trained models result in lower perplexity (better fitness).

the topic graph with thousands of topics on millions of documents. The basic idea of scalable CTM is to introduce a set of auxiliary variables, known as Polya-Gamma variables, and transform the non-conjugacy into conditional conjugacy, and thereby a partially collapsed Gibbs sampler can be developed for a distributed cluster.

Scalable CTM + postprocessing. Although scalable CTM works pretty well for most corpora, it may produce imperfect correlation results, especially for short documents like tweets. To remedy this, we utilize metadata such as retweet, co-hashtag, and the same user ratio in different topics to refine the correlation structures learned by scalable CTM. For example, if the tweets in two topics often retweet each other, the two topics are likely to be connected.

Hierarchy building. To support the navigation of large graphs, we build hierarchies for topic graphs based on the Bayesian Rose Tree (BRT) model [6, 33]. To make sure that the hierarchies built for different graphs have similar structures, we utilize the constrained BRT algorithm [52], which generates a hierarchy for each graph and iteratively refines each hierarchy by regarding the hierarchies of the other graphs as constraints.

8 Evaluation

In this section, we conduct a quantitative evaluation and a case study to demonstrate the usefulness and effectiveness of TopicPanorama. We also interview domain experts to collect their feedback.

8.1 Quantitative Evaluation

We conduct two experiments on a workstation with an Intel Xeon E5620 CPU (2.4 GHz) and 12GB Memory to demonstrate the effectiveness of the graph matching algorithm.

The first experiment shows why unified CTM with the same parameters does not work well for all corpora. Three datasets are used. Dataset A is collected from Boardreader [1] (from Jul. 2008 to Apr. 2009). It contains a news corpus (26,538 news articles), a blog corpus (13,424 blogs), and a BBS corpus (15,272 posts). Dataset B includes news articles related to Google (147,887), Microsoft (100,134), and Yahoo (14,978). Dataset C consists of Chinese news articles related to Baidu (16,723), Alibaba (12,925), and Tencent (39,074). For each dataset, we learn four topic graphs with 100 topics using CTM. Three of the topic graphs are learned separately by using the documents in each corpus, and the last topic graph was learned by using all documents in the three corpora. Table 1 shows how each graph fits the documents in each corpus. Here we utilize perplexity to measure how well the CTM model predicts a sample. Lower perplexity indicates better fitness of the topic graph with the actual data. The empirical results demonstrate that separately learned graphs have better perplexity than the jointly learned graph in each dataset. The results also imply that a consistent graph matching algorithm is needed to match multiple graphs learned from different corpora in order to develop a fuller picture of relevant topics.

(a) Dataset A

	$|\mathcal{D}|$	$|\mathcal{V}|$	$|\mathcal{E}|$
News	26,538	60	68
Blog	13,424	50	51
BBS	15,272	59	86

(b) Dataset D

	$|\mathcal{D}|$	$|\mathcal{V}|$	$|\mathcal{E}|$
Google	54,338	93	152
Microsoft	37,001	115	230
Yahoo	10,701	112	176

Table 2. Summary statistics of two datasets. $|\mathcal{D}|$ is the document number. $|\mathcal{V}|$ and $|\mathcal{E}|$ denote the node number and edge number in a topic graph.

In the second experiment, we compare the performance of our graph matching method with two baseline methods (Sec. 4.1) in terms of precision, recall, and consistency. Two human labeled datasets are used in this experiment. Two PhD students majored in text mining and familiar with the datasets labeled the matching results and the inter-annotator agreement is 87.3%. The first dataset is Dataset A, used in the first experiment. To reduce labeling efforts, the second dataset is a sampling of Dataset B. Table 2 shows the summary statistics of the two datasets.

	Dataset A				Dataset D			
	Prec.	Rec.	Conf.	Time	Prec.	Rec.	Conf.	Time
Ours	**0.81**	**0.79**	**0**	1.3	**0.79**	**0.92**	**0**	8.8
Baseline 1	0.79	0.77	4	1.2	0.69	0.85	10	8.5
Baseline 2	0.77	0.67	**0**	**0.8**	0.69	0.76	**0**	**5.7**

Table 3. Comparison of our method with the baseline methods in terms of precision (Prec.), recall (Rec.), number of conflicts (Conf.), and running time (Time) in seconds.

As shown in Table 3, our method performs better than the baseline methods with respect to precision, recall, and the number of conflicts. The precision and recall of Baseline 1 is comparable to that of our method. However, the matching result contains some conflicts. Baseline 2 can generate consistent matching, but has the lowest level of precision and recall. Moreover, the time cost of our method is comparable to that of the two baseline methods.

8.2 Case Study

We have worked closely with domain experts to develop scenarios and conduct case studies. Due to the page limit, we report one of them.

This case study aims to illustrate how TopicPanorama helps analysts meet their analytical needs and point out what functions are useful for performing related tasks. It also demonstrates the capability of TopicPanorama in handling big data from different sources and in varying formats. Two datasets are utilized, Dataset B and a Twitter dataset

(a) Dataset B (News)

	$\lvert\mathcal{D}\rvert$	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$
G	147,887	260	713
M	100,134	314	1285
Y	6,280	246	872

(b) Dataset E (Twitter)

	$\lvert\mathcal{D}\rvert$	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$
G	1,312,440	390	2292
M	2,249,610	310	1883
Y	1,588,941	370	2082

Table 4. Summary statistics of the two datasets used in the case study, one news and one Twitter corpus. G: Google; M: Microsoft; Y: Yahoo.

(Dataset D) that contains 5,150,991 tweets related to three IT companies, Google, Microsoft and Yahoo (from Jan. 2013 to Dec. 2013). Table 4 summarizes the statistics of the datasets. One expert, who has been a public relations (PR) manager for over 10 years, participated in the case study. She successfully used TopicPanorama to find a set of patterns within 2 hours and with some minor guidance from us.

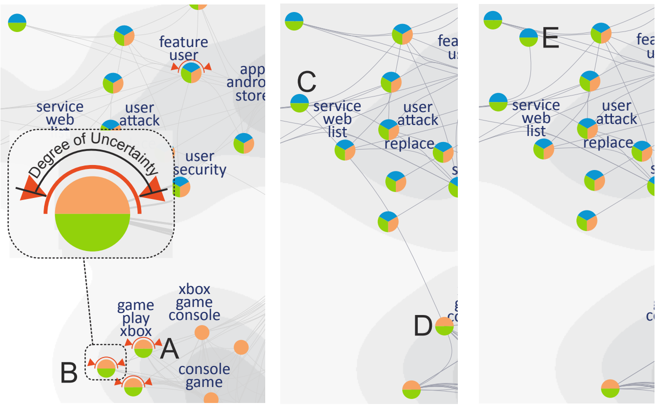
We first provide the PR manager with a full picture of three topic graphs learned from the news articles related to the three companies (Fig. 1). From the overview (Fig. 1(a)), the PR manager immediately identifies several common topics and distinctive topics of each corpus. For example, search and market related topics are shared by three corpora (Fig. 1A). Most phone related topics are shared between Google and Microsoft and a few of them are shared by the three companies (Fig. 1B). Some government related topics are referred to by the three companies and some of them are shared between Google and Yahoo (Fig. 1C). Car related topics are mainly discussed in the Google corpus (Fig. 1D). Kinect related topics are most often mentioned in the Microsoft corpus (Fig. 1E). The Yahoo corpus has some distinctive topics related to its CEO, Marissa Mayer (Fig. 1F).

The expert wanted to understand why so many government-related topics were shared by these companies. She zoomed into the fourth level of the topic tree by selecting the largest common tree node each time. As shown in Fig. 1(b), the corresponding topics can be grouped into three categories, NSA Prism spying scandal shared by the three companies (Fig. 1G), NSA Prism spying scandal shared by Google and Yahoo (Fig. 1H), and government related legal issues (Fig. 1I).

She further explored the common topics in Fig. 1G, which is enlarged in Fig. 1(d). The four topics were classified into two groups. The first group is about the disclosure of the scandal (Fig. 1G). For example, one news article was titled “NSA, FBI secretly mines data from major Internet companies.” The second category talks about actions taken by the three companies (Fig. 1K, Fig. 1L, and Fig. 1M), specifically, how they responded to this scandal in a similar manner. First, they denied cooperation with the government in disclosing users’ data (Fig. 1K). Google and Microsoft published transparency reports one after another, to disclose information about secret government requests for data. Later, Yahoo also disclosed the data requests from the US government. Second, the three companies encrypted information flowing between its various data centers (Fig. 1L). In this action, Google took the lead, with Yahoo responding similarly, and Microsoft later joining Google and Yahoo in beefing up encryption. The expert originally believed that only Google and Yahoo encrypted their data centers. After exploring the related topics with our tool, she found that Microsoft also stepped up encryption to thwart the NSA. She commented, “This is a surprise to me. I really appreciate this tool because it corrects my wrong understanding.” Finally, the three companies and other major tech companies asked the US government to reform surveillance laws (Fig. 1M).

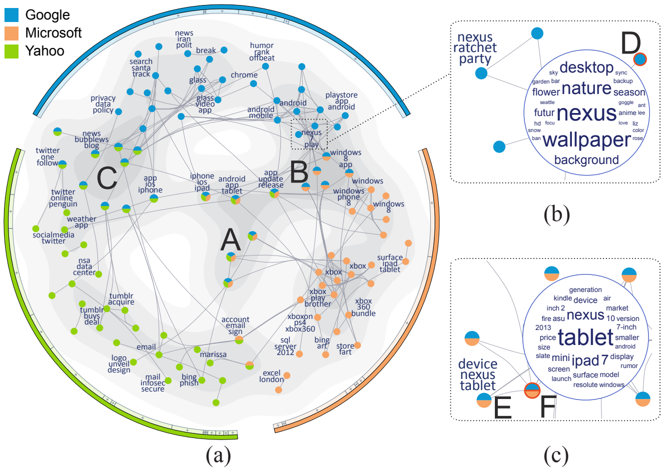
In the above exploration, the expert found one interesting pattern. When publishing the reports, Yahoo followed Google and Microsoft. However, Yahoo was more active in making plans to encrypt information. The expert was curious about such a change, so she continued to explore the topics correlated to both topics shown in Fig. 1K and Fig. 1L. After some exploration, she found a relevant topic that talked about “NSA statement on Washington Post report on infiltration of Google, Yahoo data center links” (Fig. 1N), which was connected to each of these two topics, respectively.

The expert was interested in game related topics, so she entered “game” into the search box. The search result is shown in Fig. 10(a). She enabled the tool to show the uncertainty glyph of the matched topics. After examining the results, she found two incorrect mappings, **A** and **B**, which map Microsoft Xbox games to Yahoo sports related games. She first unmatched **A**. In the new mapping result (Fig. 10(b)), **A** was changed to **C**. **B** was changed to **D**, which is still an incorrect mapping. She then unmatched **D**, which was changed to **E** in Fig. 10(c).



(a) (b) (c)

Figure 10. Interactive editing of the graph matching result.

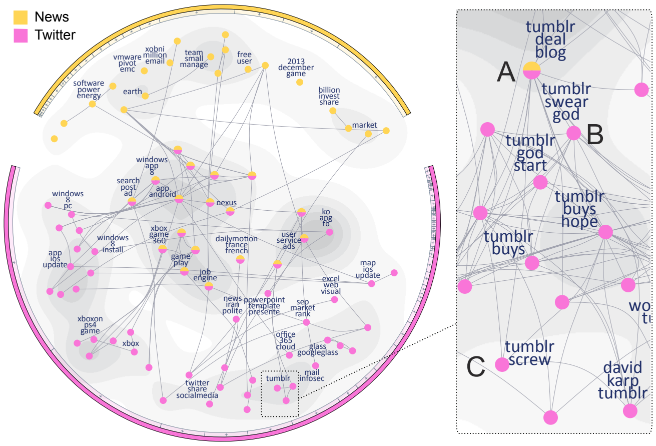


(a) (b) (c)

Figure 11. The full picture of Google, Microsoft, and Yahoo in the Twitter corpus. (a) Overview of the Twitter dataset; (b) Nexus related topics in Twitter; (c) Nexus related topics in news.

In addition to news topics, the expert was eager to find the corresponding topics on Twitter. Thus we provided a full picture of all tweets related to the three companies (Fig. 11(a)). By looking at the overview, she immediately observed that the Twitter topics were less correlated compared with the news topics. Moreover, the number of common topics is less than that of news (Fig. 11A, Fig. 11B, and Fig. 11C). For example, for topics related to Nexus, most Twitter topics are unique to the Google corpus (Fig. 11(b)) while most news topics are common ones between Google and Microsoft (Fig. 11(c)). The Twitter topics focused more on specific features of Nexus (Fig. 11D). While the news topics talked about the launch of Nexus (“Google to launch new Nexus 7 tablet in July for $229: Report”, Fig. 11E) and comparison with similar products (“New Nexus 7 vs iPad Mini. Screen Resolution Price and Specs”, Fig. 11F).

To better compare topics between news and Twitter, we output the two matched graphs as two single graphs and then matched them together. The expert observed that there were more Twitter topics than news topics (Fig. 12(a)). After some exploration, the Tumblr related topics attracted her attention. Among these topics, there was only one common topic and the rest of the topics were from Twitter. The common topic was about the acquisition of Tumblr (Fig. 12A). The unique topics focused on giving opinions such as “this whole yahoo and Tumblr relationship is painful. I don’t want it” (Fig. 12B) and providing information or suggestions such as “Three Ways Yahoo Can Avoid Screwing Up Tumblr.” (Fig. 12C) After studying these Twitter topics, the expert said, “It is good to know there are so many



(a) (b)

Figure 12. Matching the news corpus with the Twitter corpus: (a) overview; (b) comparison of Tumblr related topics.

complaints about the acquisition, which will help the company to take corresponding actions."

8.3 Expert Interview

We have interviewed the six domain experts working with us and verified interview transcripts with them. We conducted a semi-structured interview guided by a predefined questionnaire that contains a list of usability and effectiveness related questions. Each of the evaluations took around 90 minutes, including 10 minutes of system introduction, 50 minutes of case study and free exploration, and 30 minutes for the post interview. Overall, TopicPanorama has been well received by the experts. The feedback is summarized into four themes.

Graph matching. All the experts agreed that the graph matching component is very useful for developing a full picture of relevant topics that are discussed across multiple sources. They especially liked the interactive editing function.

Interactive visualization. The experts were impressed by the power of the visualization components. They all liked the hybrid visualization that allows them to understand the full picture at different granularity levels. They strongly agreed that the node-link diagram enables them to find topics of interest quickly and the stacked tree allows them to smoothly drill into detailed cluster layers. Furthermore, the uncertainty glyph provides an easy way to examine the mapping results with lower scores. The experts can then freely modify the error matching result based on their knowledge.

Insight discovery. All the experts were able to use TopicPanorama to form a full picture of relevant topics across multiple sources. They were able to find topics of interest and then drill down to examine their relationships with other topics. With TopicPanorama, experts were even able to gain insight they did not have before. For example, a senior public relations manager of a large IT company believed that in the NSA Prism spying scandal, only Google and Yahoo encrypted data while Microsoft did not. Based on what she saw by exploring the related topics in our tool, she learned that Microsoft also beefed up encryption following the actions of the other two companies. All the experts were intrigued by the TopicPanorama application. They believe TopicPanorama can better help with their analysis. For example, one sociologist said, "I am eager to apply this tool to the competitor analysis project I conducted for three TV networks in the US: ABC, CBS, and NBC. I want to develop a full picture of the competition graph of audience loyalty of the three TV networks. Specifically, I want to have a full picture that illustrates who are the most loyalty audiences of each network and who switch back and forth between the three networks."

Improvements. The experts also suggested several improvements. Four of the experts expressed the need to analyze temporal patterns of the matched graph because understanding such patterns and identifying the major causes leading to them are very important in their work. Two experts wanted to add some domain-specific information to the graph matching process, for example, allowing the user to explicitly specify which keywords do not contribute to the node mappings, with TopicPanorama incrementally updating the mapping results to reflect such user requirements.

9 Conclusions and Future Work

We have worked closely with a group of experts, including public relations managers, journalists, and sociologists, to derive several high-level tasks. Based on these tasks, we developed TopicPanorama to help users develop a full picture of relevant topics that are discussed in multiple sources. In close collaboration with domain experts at every stage of development, we have iteratively refined and improved the toolkit, including the mining and visualization components.

The system provides three advantages over existing techniques. First, it derives consistent graph mapping results among multiple graphs efficiently. Second, it provides an LOD-based visualization that allows the user to examine the mapping results globally and locally and switch between the global overview and local details smoothly. Third, it allows the user to incrementally edit the mapping results according to their knowledge and information needs.

Our design also has some limitations. Although our graph matching algorithm and visualization method can handle any number of graphs, the number of corpora that can be visually compared is not large due to visual clutter and limited screen space. According to the interview with experts, they can leverage TopicPanorama to analyze two or three topic graph mapping results very well. It also works for four topic graphs though it takes longer to gain insight. It may fail to provide a better understanding for five or more topic graphs due to the limited display area and complex mapping results. Previous experiments [26, 60] have consistently found that approximately four objects can be tracked in visual comparison. This conclusion is consistent with the feedback of our target users. They said they usually work on two or three corpora and seldom analyze four corpora in their work. Consequently, TopicPanorama works for most real-world applications. Another limitation is not all the topics in the topic graph are meaningful. In our current implementation, we rank the topics and filter the less important ones. A possible solution is to allow users to interactively edit topic mining results [12].

Future work will include the extension of interactive editing of mapping results to topic mining results. The key is to study how to effectively combine the topic mining model with our graph matching algorithm. Another exciting avenue for future work is to design a suitable visualization for more than three corpora. Additionally, we are interested in analyzing temporal evolution patterns of the common and distinctive topics in the matched graph. We would also like to study how to efficiently integrate domain-specific information such as the relationships between keywords and node mappings into the incremental graph matching algorithm.

Acknowledgments

The authors would like to thank K. Zhou and J. Yang for the insightful discussions and S. Lin for proofreading the paper. J. Chen and J. Zhu are supported by the National Basic Research Program of China (No. 2013CB329403), National Natural Science Foundation of China (No.s 61322308, 61332007), a Microsoft Research Fund (No. FY14-RES-SPONSOR-111), and the National University Student Innovation Program.

References

[1] Boardreader. `http://www.boardreader.com`, Mar. 2014.

[2] B. Alper, B. Bach, N. H. Riche, T. Isenberg, and J.-D. Fekete. Weighted graph comparison techniques for brain connectivity analysis. In *CHI*, pages 483–492, 2013.

[3] K. Andrews, M. Wohlfahrt, and G. Wurzinger. Visual graph comparison. In *IV*, pages 62–67, 2009.

[4] D. Blei and J. Lafferty. Correlated topic models. In *NIPS*, pages 147–154. 2006.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[6] C. Blundell, Y. W. Teh, and K. A. Heller. Bayesian rose trees. In *UAI*, pages 65–72, 2010.

[7] J. Branke. Dynamic graph drawing. In *Drawing Graphs*, pages 228–246, 1999.

[8] S. Bremm, T. von Landesberger, M. Hess, T. Schreck, P. Weil, and K. Hamacher. Interactive visual comparison of multiple trees. In *IEEE VAST*, pages 31–40, 2011.

[9] N. Cao, J. Sun, Y.-R. Lin, D. Gotz, S. Liu, and H. Qu. Facetatlas: Multifaceted visualization for rich text corpora. *IEEE TVCG*, 16(6):1172–1181, 2010.

[10] A. J.-B. Chaney and D. M. Blei. Visualizing topic models. In *ICWSM*, 2012.

[11] J. Chen, J. Zhu, Z. Wang, X. Zheng, and B. Zhang. Scalable inference for logistic-normal topic models. In *NIPS*, pages 2445–2453. 2013.

[12] J. Choo, C. Lee, C. K. Reddy, and H. Park. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE TVCG*, 19(12):1992–2001, 2013.

[13] C. Collins and M. S. T. Carpendale. Vislink: Revealing relationships amongst visualizations. *IEEE TVCG*, 13(6):1192–1199, 2007.

[14] D. Conte, P. Foggia, C. Sansone, and M. Vento. Thirty years of graph matching in pattern recognition. *International journal of pattern recognition and artificial intelligence*, 18(03):265–298, 2004.

[15] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong. Textflow: Towards better understanding of evolving topics in text. *IEEE TVCG*, 17(12):2412–2421, 2011.

[16] W. Cui, S. Liu, Z. Wu, and H. Wei. How hierarchical topics evolve in large text corpora. *To appear in IEEE TVCG*, 2014.

[17] W. Cui, X. Wang, S. Liu, N. H. Riche, T. M. Madhyastha, K. L. Ma, and B. Guo. Let it flow: a static method for exploring dynamic graphs. In *IEEE PacificVis*, pages 121–128, 2014.

[18] M. Dörk, D. M. Gruen, C. Williamson, and M. S. T. Carpendale. A visual backchannel for large-scale events. *IEEE TVCG*, 16(6):1129–1138, 2010.

[19] W. Dou, X. Wang, R. Chang, and W. Ribarsky. Paralleltopics: A probabilistic approach to exploring document collections. In *IEEE VAST*, pages 231–240. IEEE, 2011.

[20] W. Dou, L. Yu, X. Wang, Z. Ma, and W. Ribarsky. Hierarchicaltopics: Visually exploring large text collections using topic hierarchies. *IEEE TVCG*, 19(12):2002–2011, 2013.

[21] X. Gao, B. Xiao, D. Tao, and X. Li. A survey of graph edit distance. *Pattern Analysis and applications*, 13(1):113–129, 2010.

[22] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, 2011.

[23] M. Hascoët and P. Dragicevic. Interactive graph matching and visual comparison of graphs and clustered graphs. In *AVI*, pages 522–529, 2012.

[24] S. Havre, E. G. Hetzler, P. Whitney, and L. T. Nowell. Themeriver: visualizing thematic changes in large document collections. *IEEE TVCG*, 8(1):9–20, 2002.

[25] N. Henry, J. Fekete, and M. J. McGuffin. Nodetrix: a hybrid visualization of social networks. *IEEE TVCG*, 13(6):1302–1309, 2007.

[26] J. Intriligator and P. Cavanagh. The spatial resolution of visual attention. *Cognitive psychology*, 43(3):171–216, 2001.

[27] T. Kamada and S. Kawai. An algorithm for drawing general undirected graphs. *Information processing letters*, 31(1):7–15, 1989.

[28] G. A. Korsah, A. T. Stentz, and M. B. Dias. The dynamic hungarian algorithm for the assignment problem with changing costs. Technical Report CMU-RI-TR-07-27, July 2007.

[29] G. Kumar and M. Garland. Visual exploration of complex time-varying graphs. *IEEE TVCG*, 12(5):805–812, 2006.

[30] O. D. Lampe and H. Hauser. Interactive visualization of streaming data with kernel density estimation. In *PacificVis*, pages 171–178, 2011.

[31] S. Liu, W. Cui, Y. Wu, and M. Liu. A survey on information visualization: recent advances and challenges. *The Visual Computer*, pages 1–21, 2014.

[32] S. Liu, M. X. Zhou, S. Pan, Y. Song, W. Qian, W. Cai, and X. Lian. Tiara: Interactive, topic-based visual text summarization and analysis. *ACM TIST*, 3(2):25, 2012.

[33] X. Liu, Y. Song, S. Liu, and H. Wang. Automatic taxonomy construction from keywords. In *KDD*, pages 1433–1441, 2012.

[34] A. M. MacEachren, R. E. Roth, J. O'Brien, B. Li, D. Swingley, and M. Gahegan. Visual semiotics & uncertainty visualization: An empirical study. *IEEE TVCG*, 18(12):2496–2505, 2012.

[35] P. McLachlan, T. Munzner, E. Koutsofios, and S. North. Liverac: interactive visual exploration of system management time-series data. In *CHI*, pages 1483–1492, 2008.

[36] J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial & Applied Mathematics*, 5(1):32–38, 1957.

[37] T. Munzner. A nested process model for visualization design and validation. *IEEE TVCG*, 15(6):921–928, 2009.

[38] T. Munzner, F. Guimbretière, S. Tasiran, L. Zhang, and Y. Zhou. Treejuxtaposer: scalable tree comparison using focus+ context with guaranteed visibility. In *ACM TOG*, volume 22, pages 453–462, 2003.

[39] R. Myers, R. Wison, and E. R. Hancock. Bayesian graph edit distance. *IEEE PAMI*, 22(6):628–635, 2000.

[40] M. Neuhaus and H. Bunke. A convolution edit kernel for error-tolerant graph matching. In *ICPR*, volume 4, pages 220–223, 2006.

[41] A. Ribalta and F. Serratosa. On the computation of the common labelling of a set of attributed graphs. In *CIARP*, pages 137–144, 2009.

[42] A. Ribalta and F. Serratosa. Models and algorithms for computing the common labelling of a set of attributed graphs. *Computer Vision and Image Understanding*, 115(7):929–945, 2011.

[43] K. Riesen and H. Bunke. Approximate graph edit distance computation by means of bipartite graph matching. *Image Vision Comput.*, 27(7):950–959, 2009.

[44] K. Riesen, X. Jiang, and H. Bunke. Exact and inexact graph matching: Methodology and applications. In *Managing and Mining Graph Data*, pages 217–247. 2010.

[45] K. Salomatin, Y. Yang, and A. Lad. Multi-field correlated topic modeling. In *SDM*, pages 628–637, 2009.

[46] R. R. Sambasivan, I. Shafer, M. L. Mazurek, and G. R. Ganger. Visualizing request-flow comparison to aid performance diagnosis in distributed systems. *IEEE TVCG*, 19(12):2466–2475, 2013.

[47] P. N. Suganthan, E. K. Teoh, and D. P. Mital. Pattern recognition by graph matching using the potts mft neural networks. *Pattern Recognition*, 28(7):997–1009, 1995.

[48] G. Sun, Y. Wu, R. Liang, and S. Liu. A survey of visual analytics techniques and applications: State-of-the-art research and future challenges. *Journal of Computer Science and Technology*, 28(5):852–867, 2013.

[49] G. Sun, Y. Wu, S. Liu, T.-Q. Peng, J. J. H. Zhu, and R. Liang. EvoRiver: Visual analysis of topic coopetition on social media. *To appear in IEEE TVCG*, 2014.

[50] C. Vehlow, T. Reinhardt, and D. Weiskopf. Visualizing fuzzy overlapping communities in networks. *IEEE TVCG*, 19(12):2486–2495, 2013.

[51] T. Von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J. J. van Wijk, J.-D. Fekete, and D. W. Fellner. Visual analysis of large graphs: State-of-the-art and future research challenges. *Computer graphics forum*, 30(6):1719–1749, 2011.

[52] X. Wang, S. Liu, Y. Song, and B. Guo. Mining evolutionary multi-branch trees from text streams. In *KDD*, pages 722–730, 2013.

[53] M. L. Williams, R. C. Wilson, and E. R. Hancock. Multiple graph matching with bayesian inference. *Pattern Recognition Letters*, 18(11-13):1275–128, 1997.

[54] R. C. Wilson and E. R. Hancock. Structural matching by discrete relaxation. *IEEE PAMI*, 19(6):634–648, 1997.

[55] R. C. Wilson, E. R. Hancock, and B. Luo. Pattern vectors from algebraic graph theory. *IEEE PAMI*, 27(7):1112–1124, 2005.

[56] Y. Wu, F. Wei, S. Liu, N. Au, W. Cui, H. Zhou, and H. Qu. Opinionseer: Interactive visualization of hotel customer feedback. *IEEE TVCG*, 16(6):1109–1118, 2010.

[57] Y. Wu, G.-X. Yuan, and K.-L. Ma. Visualizing flow of uncertainty through analytical processes. *IEEE TVCG*, 18(12):2526–2535, 2012.

[58] P. Xu, Y. Wu, E. Wei, T.-Q. Peng, S. Liu, J. J. H. Zhu, and H. Qu. Visual analysis of topic competition on social media. *IEEE TVCG*, 19(12):2012–2021, 2013.

[59] J. Yan, Y. Tian, H. Zha, X. Yang, Y. Zhang, and S. M. Chu. Joint optimization for consistent multiple graph matching. In *ICCV*, pages 1649–1656, 2013.

[60] S. Yantis. Multielement visual tracking: Attention and perceptual organization. *Cognitive psychology*, 24(3):295–340, 1992.

[61] J. Zhang, Y. Song, C. Zhang, and S. Liu. Evolutionary hierarchical dirichlet processes for multiple correlated time-varying corpora. In *KDD*, pages 1079–1088, 2010.