EuroVis Workshop on Visual Analytics (2018)
C. Tominski and T. von Landesberger (Editors)

ComModeler: Topic Modeling Using Community Detection

Tommy Dang and Vinh T. Nguyen

Computer Science Department, Texas Tech University, Lubbock, USA

Abstract

This paper introduces ComModeler, a novel approach for topic modeling using community finding in dynamic networks. Our algorithm first extracts the terms/keywords, formulates a network of collocated terms, then refines the network based on various features (such as term/relationship frequency, sudden changes in their frequency time series, or vertex betweenness centrality) to reveal the structures/communities in dynamic social networks. These communities correspond to different hidden topics in the input text documents. Although initially motivated to analyze text documents, we soon realized the ComModeler has more general implications for other application domains. We demonstrate the ComModeler on several real-world datasets, including the IEEE VIS publications from 1990 to 2016, together with collocated phrases obtained from various political blogs.

1. Introduction

In the field of machine learning, topic modeling such as Latent Dirichlet allocation (LDA) [BNJ03] has emerged as a tool to unveil hidden patterns from a large corpus of documents that best represents the information in the collection. Topic modeling enables us to summarize the text corpora, and hence it is highly used for the visualization of the large data in a time efficient manner [WM06, WLS*10]. However, topic model itself cannot disclose hot topics and their coherent relationships in text corpora [MCZZ08]. We approach this problem from a very different perspective by building a network based on text similarities and utilizing the network features such as *node degree* and *betweenness centrality*. Communities extracted from these refined networks serve as popular topics at each time point. Our contributions in this paper thus are:

- We propose a new approach to discover topics using community detection in networks. The approach allows us to refine the network and reveal its community formations.
- We develop an interactive prototype, named *ComModeler*, to explore and visualize topic abstractions. The *ComModeler* supports a range of interactive features, such as lensing and filtering, allowing users to narrow down events of interest quickly.
- We demonstrate the *ComModeler* on various text corpus of news, such as *Wikinews* and political blogs and provide use cases of *ComModeler* in other application domains.

The rest of this paper is organized as follows. Section 2 reviews related work and existing methodologies. Section 3 introduces our *ComModeler* prototype and illustrates it on news datasets. In Section 4, we argue that although originally motivated to deal with text corpora, the *ComModeler* has applications to many other domains as well. Finally, we conclude our paper with future work.

2. Related Work

Existing content-based community detection methods consider the social media users to be related when they share common interests and use this measure to group the users [ZLZG13,DOG15]. Fani et al. [FZBD16] improve the quality of detected communities by taking into account the users' temporal behavior towards their topics of interest to identify like-minded users. Topic over Time [WM06] model that jointly captures keyword co-occurrences and locality of those patterns over time is utilized to discover users' topics of interest. This paper focuses on detecting temporal events concealed in large text corpora, not the users' social networks. Moreover, our interactive prototype allows users to visually explore, customize, and make discoveries by themselves. The relations between entities play an important role in interpreting the dynamic of topics. Instead of using the notion of a "bag of words" as in LDA, we approach the problem by taking the advantage of the network properties. The community detection aims to partition a network into clusters such that entities are strongly connected within their cluster and weakly connected with other clusters.

Network modularity: Modularity [New06] is an objective function to measure the quality of the communities in a network. Modularity is defined as the fraction of the edges that fall within the given groups minus the expected fraction if edges were distributed at random. A network with high modularity is indicating that it has dense connections between nodes within a community and has sparse connections between nodes of different communities. Modularity Q measure extends from 0 to 1.

The Louvain method: One of the most widely used algorithms for community detection in practice is the Louvain method [BGLL08] because of its speed and the quality of graph partitions [RN11]. It is a greedy optimization method that initially finds small communities by locally maximizing the modularity and

DOI: 10.2312/eurova.20181104

consequently performs the same procedure on the new network by considering each community extracted in the previous step as a single vertex. This procedure iterates until a maximum of modularity Q is obtained, and a hierarchy of communities is produced.

Betweenness centrality: To track the importance of vertices, we use betweenness centrality [Bra01].There are many centrality measures such as *closeness centralization*, *eigenvector centrality*, and *PageRank* [PBMW99], a eigenvector centrality which is used by Google Search to rank websites in their search engine results. We select *degree centrality* and *betweenness centrality* to track node importance (in Section 3.3) since they often provide a useful means of identifying bridges and brokers, such as influential people in a social network.

3. *ComModeler* Visualization

This paper proposes an approach applying a graph partitioning algorithm to extract clusters of important terms/phrases that form latent communities to uncover social topics. *ComModeler* aims to achieve the following high-level goals [Mun09, MSM12, BM13] when analyzing text corpus:

Detect communities based on network coherence. To extract entities' relationship from a dataset, we construct semantic networks from various news and discussions: A node could be people, an organization, or a city, and two nodes are connected whenever they appear in the same contexts. Edges' weights are calculated based on mention frequency of the entities at each time unit. State-of-the-art community detection algorithms are applied to automatically group related entities in a network snapshot.

Identify influential people or events. Identifying the most influential people or events from huge networks is extremely challenging. We tackle this problem by constructing collocation-based relationship networks, we present each node's frequency, degree, and betweenness centrality in a detailed measure. Based on these measurements, the *ComModeler* filters out trivial nodes or edges and preserves sub-graphs which meet filtering conditions. In this way, the *ComModeler* is able to pinpoint the most influential people or events, and show the results in corresponding word clouds.

Discover interesting topics across the network. Topic modeling can provide an overall landscape of topic distribution in documents. As LDA uses a "bag of words" approach and treats each document as a vector of word counts, it has the limitation of providing interactions and topic connections. Many applications are still faced challenges such as choosing a suitable number of topics and providing Bayesian justification. Our approach incorporates the advantages of network properties, which can trace back to historical topics when a similar topic appears again, pinpoint the original timestamps, and connect with high co-allocation terms, thus overcomes the limitations of the LDA model by providing more interactive choices.

Discern major topic changes over time. Besides the topics displayed for each network, the *ComModeler* reviews the dynamic process of topics becoming popular or cooling down over time. Most interestingly, when an old topic is mentioned again, the *ComModeler* will trace back and point out its latest die out time.

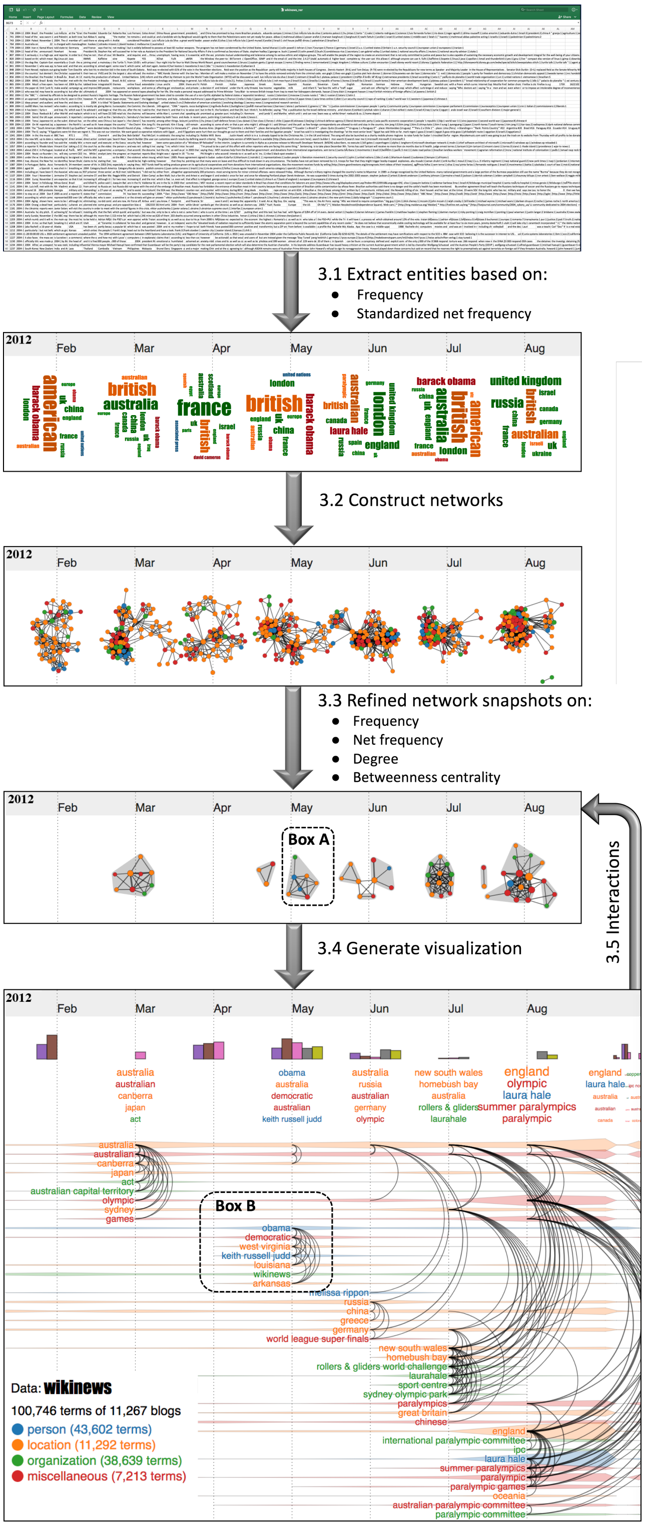

Figure 1: *A schematic overview showing the main components of ComModeler: Extracting popular terms, constructing the relationships of ranked terms, filtering the network snapshots based on different node/link properties, visualizing the dynamic of extracted topics, and interacting with the visualization to explore the events of interests. The data is retrieved from Wikinews articles in 2012.*

Our prototype focus on four low-level analysis tasks:

- **T1**: Provide a summary view of text corpus over time [KPS04]. Our prototype provides a quick overview of important topics using the network thumbnails. Moreover, we display a summary histogram of network modularity for different settings as well as the top 5 popular terms at each time unit. (Section 3.2)
- **T2**: Retrieve and display network details on demand. Users can expand several consecutive windows by moving the lens over timeline on the top. (Section 3.3).
- **T3**: Filter terms/topics on user request. For example, users may want to see political events in a specific geographic area or other discussed topics/activities related to a particular person.
- **T4**: Sort terms in the text clouds based on a user selected measure: *term frequency*, *sudden increase in frequency*, *vertex degree*, or *betweenness centrality* (Section 3.4). Ordering is also applied to time series graph to reduce visual clutter.

Figure 1 shows a schematic overview of *ComModeler* visualization. In this section, we describe *ComModeler* fundamental components w.r.t. topic modeling in text corpus. However, *ComModeler* has more general applications which will be discussed in Sections 4.2. To enable users to explore the vast temporal text corpora in an interactive and efficient way, *ComModeler* adopts the following steps and components:

3.1. Extract terms

The input text documents are preprocessed into entities and further classified into different categories [Mon17]. A common way to represent term frequencies is using word clouds (one for each time stamp) as depicted in the second panel of Figure 1. A limitation of this technique (side-by-side snapshot techniques in general) is when the number of entities/terms increase, keeping track of individual entities and comparing multiple snapshots requires more mental efforts [BBDW16].

3.2. Construct networks for each time point

There are several methods to generate relationships among two given text elements. For instance, a link can be generated by the similarity between two documents [JKS*14], the mentioned frequencies in the same contexts [DPF16], or the defined logical forms [KD08]. Since *ComModeler* works directly on individual terms to obtain community-based coherence, the relationship is determined based on the co-occurrences of the terms in the same articles/blogs. The third panel of Figure 1 shows an example of the force layouts of terms extracted from *Wikinews* in 7 months from February to August 2012. The thickness of a link connecting two terms A and B represents how often A and B are collocated together in the same news articles for a given time unit, such as a week or a month (which can be decided by the user). The force layouts are implemented in D3.js [BOH11] on separated Scalable Vector Graphics (SVG) which can be zoomed in/out (when lensing is applied) by simply changing the SVGs' *viewBox*.

3.3. Refine the network snapshots

The *ComModeler* allows users to narrow down the text network using different parameters (visualization task **T3**). In particular, users can refine the network snapshots using: non-network properties (including *term frequency* and *sudden increase in frequency*) and vertex properties (including *vertex degree* and *betweenness centrality*). While there are many possible centrality measures such as *closeness centralization* and *eigenvector centrality*, we select *vertex degree* and *betweenness centrality* since they provides effective means of identifying bridges and brokers.

The *ComModeler* pre-computes and achieves the sub-networks (and their modularity scores) of various *edge weight threshold* settings by leaving the "uninteresting" elements to be filtered and reconsidered later. At first, the refined sub-networks which have highest cluster separations (or max modularity Q generated by the Louvain method [BGLL08]) at each time point are presented. The edge weight filtering operation is then offered to the user. The fourth panel of Figure 1 shows an example of refining the previous network by filtering (on the user's demand) only terms that mentioned at least five times together in one month in *Wikinews* articles. Notice that there are some empty networks (February and April 2012) since there is no connections between terms satisfied the filtering condition: link weight (or collocated frequency) is at least 5. The detected cluster in Box A is about the event on May 8, 2012, Keith Russell Judd won 41% of the primary vote in West Virginia (Democratic primary) against incumbent Barack Obama, a higher percentage of the vote in one state than any other primary opponent of Obama had hitherto achieved in 2012. These community details are also highlighted in Box B in the last panel of Figure 1.

3.4. Generate visualizations

CloudLines style visualization [KBK11, LYK*12] are popular for representing the evolution of terms over time. The drawback of this approach is that it lacks explicit relationship representations. To mitigate this issue, arcs diagrams [GBD09] are overlaid on top of time series to highlight correlation of terms. Small multiples (one for each term) are ordered vertically by their time span. In particular, some terms which are often mentioned in later months can be easily noticed by larger arcs from a lower term group (of the later month) to an upper group (of previous months) as depicted at the last panel of Figure 1. Terms in the same month are ordered by the detected community to reduce edge-crossings (visualization task **T4**) since the community detection algorithm groups highly connected nodes (frequently collocated terms) while separate loosely connected nodes into different clusters. The quality of the produced cluster formation is reflected in modularity presented in the histograms which are color-coded by different filtering conditions on term co-occurrence frequency as explained in Section 3.3. Modularity Q extends from 0 to 1. Higher Q indicates better community separation.

4. Experiments

4.1. Datasets

In this section, we illustrate the features of *ComModeler* mainly through examples. We use text datasets retrieved from different political blogs and news, IEEE VIS co-authorship networks from 1990 to 2016 to demonstrate the performance of *ComModeler*.

4.2. Use cases

Figure 2 shows an example of *ComModeler* visualization for the *EmptyWheel* dataset which contains 2,618 political blogs. In particular, we apply lensing on April to December 2013 interval; terms are filtered and ranked based on *frequency* while connections are filtered by *edge weight threshold* ≥ 3. As depicted, nodes in the network snapshots have different sizes computed based on their *frequency*. The detected communities represent important political events in each month. The *modularity* Q for each refined network snapshot is plotted in the pink bar of the histogram underneath. For example in May 2013, the pink bar (at the pink arrow) is tallest compared to other bars in the same month since the current filter (*edge weight threshold* ≥ 3) reveals two big communities in Box **D1** and Box **E1**. Box **C1** and Box **C2** highlight the *Boston marathon* bombing occurred on April 15, 2013. Two homemade bombs detonated near the finish line of the annual *Boston marathon* (red node), killing three people and injuring several hundred others. *FBI* later identified the bombers: *dzhokhar tsarnaev* and *tamerlan tsarnaev* (blue nodes). Box **D1** and Box **D2** relate to the triple homicide that took place in *Waltham*, Massachusetts. *Todashev* is a friend of *Tamerlan Tsarnaev*, the *Boston marathon* bomber. Box **F1** and Box **F2**: On June 21, 2013, the U.S. Department of Justice unsealed charges against *Edward Snowden*. Box **F3**: The connections between *Edward Snowden*, *NSA*, and other terms in this cluster indicate that this was still a hot political event in many more months after its first emergence in June 2013.

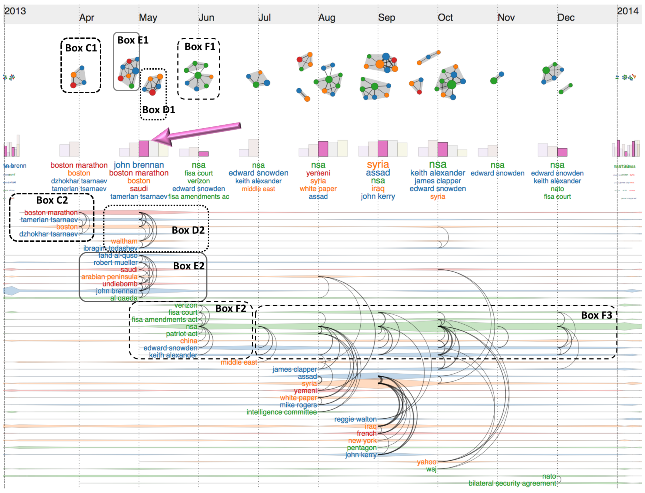

Figure 2: *ComModeler for the EmptyWheel data.*

Figure 3 shows another example of *ComModeler* visualization for the *IEEE VIS publications* data [IHK*17]. In this example, we have filtered "Kwan-Liu Ma" publications using the search box (visualization task **T3**). We lense on 2006 to June 2016 interval. Nodes are Ma's co-authors on his IEEE VIS publications which are ranked and scaled based on *publication frequency* (larger nodes are authors with more publications in a given year). Two authors are connected as they publish together to one of the IEEE VIS conferences. Authors are color-coded by their first publication venue: blue for Vis, orange for VAST, green for InfoVis, and red for SciVis.

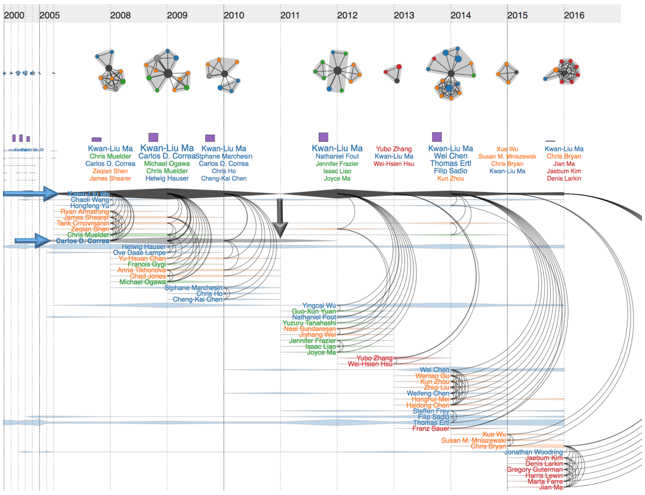

Figure 3: *ComModeler visualization for "Kwan-Liu Ma" research networks for his IEEE VIS publications from 1990 to 2016.*

The selected researcher (Kwan-Liu Ma) is highlighted in black and appears in the center of network snapshots. Communities on the tops are different groups of researchers Ma has been working with. The networks are empty in 2006 and 2011 since he did not have IEEE VIS publications in these years. The word clouds show top co-authors of Kwan-Liu Ma over time. The arcs diagrams at the bottom summarize Ma's collaboration network in the past ten years (visualization task **T1**). For example, Carlos D. Coorea (highlighted in gray in the network snapshots and located at the lower blue arrow in the time series) had continuous publications to IEEE VIS with Kwan-Liu Ma during his three years (2007-2010) working at University of California, Davis as a postdoctoral researcher. The time series at the black arrow suggest that Carlos D. Coorea has another publication in 2011 (not with Kwan-Liu Ma). Arcs diagrams are ordered by their time span. Researchers in the same year are organized by communities (or paper collaborations). As depicted at the bottom of Figure 3, this ordering strategy generates ladder-style visualization (more recent collaborations are pushed further to the right and presented in larger arcs) and hence reduces edge-crossings.

5. Conclusion

This paper proposes a novel topic modeling approach that incorporates a community detection algorithm to reveal network structure in temporal data automatically. We also introduce a visual analytics prototype which helps users to explore topic abstractions via interactive features. Our experiments with political news datasets show that *ComModeler* provide a useful lens into a large corpus of texts. Furthermore, our use cases demonstrate the usefulness of each component of the *ComModeler* on various application domains.

ComModeler is implemented in D3.js [BOH11]. The online application, source code, supplementary materials, more use cases, and a demo video are provided via our GitHub project repository, located at `https://dycomdetector.github.io/`.

References

[BBDW16] Beck F., Burch M., Diehl S., Weiskopf D.: A taxonomy and survey of dynamic graph visualization. In *Computer Graphics Forum* (2016), Wiley Online Library. 3

[BGLL08] Blondel V. D., Guillaume J.-L., Lambiotte R., Lefebvre E.: Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment 2008*, 10 (2008), P10008. 1, 3

[BM13] Brehmer M., Munzner T.: A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (Dec 2013), 2376–2385. doi:10.1109/TVCG.2013.124. 2

[BNJ03] Blei D. M., Ng A. Y., Jordan M. I.: Latent dirichlet allocation. *Journal of machine Learning research 3*, Jan (2003), 993–1022. 1

[BOH11] Bostock M., Ogievetsky V., Heer J.: D3 data-driven documents. *IEEE Trans. Vis. Comput. Graph. 17*, 12 (2011), 2301–2309. 3, 4

[Bra01] Brandes U.: A faster algorithm for betweenness centrality. *Journal of mathematical sociology 25*, 2 (2001), 163–177. 2

[DOG15] Darmon D., Omodei E., Garland J.: Followers are not enough: A multifaceted approach to community detection in online social networks. *PloS one 10*, 8 (2015), e0134860. 1

[DPF16] Dang T. N., Pendar N., Forbes A. G.: TimeArcs: Visualizing Fluctuations in Dynamic Networks. *Computer Graphics Forum* (2016). doi:10.1111/cgf.12882. 3

[FZBD16] Fani H., Zarrinkalam F., Bagheri E., Du W.: Time-sensitive topic-based communities on twitter. In *Canadian Conference on Artificial Intelligence* (2016), Springer, pp. 192–204. 1

[GBD09] Greilich M., Burch M., Diehl S.: Visualizing the evolution of compound digraphs with timearctrees. In *Proc. Eurographics Conf. on Visualization* (2009), pp. 975–990. 3

[IHK*17] Isenberg P., Heimerl F., Koch S., Isenberg T., Xu P., Stolper C. D., Sedlmair M. M., Chen J., Möller T., Stasko J.: vispubdata.org: A Metadata Collection about IEEE Visualization (VIS) Publications. *IEEE Transactions on Visualization and Computer Graphics 23* (2017). To appear. URL: https://hal.inria.fr/hal-01376597, doi:http://dx.doi.org/10.1109/TVCG.2016.2615308. 4

[JKS*14] Jin F., Khandpur R. P., Self N., Dougherty E., Guo S., Chen F., Prakash B. A., Ramakrishnan N.: Modeling mass protest adoption in social network communities using geometric brownian motion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (2014), ACM, pp. 1660–1669. 3

[KBK11] Krstajic M., Bertini E., Keim D.: CloudLines: Compact display of event episodes in multiple time-series. *IEEE Trans. Vis. Comput. Graph. 17*, 12 (2011), 2432–2439. 3

[KD08] Kok S., Domingos P.: Extracting semantic networks from text via relational clustering. *Machine Learning and Knowledge Discovery in Databases* (2008), 624–639. 3

[KPS04] Keim D. A., Panse C., Sips M.: Information visualization : Scope, techniques and opportunities for geovisualization. In *Exploring Geovisualization*, Dykes J., (Ed.). Elsevier, Oxford, 2004, pp. 1–17. 3

[LYK*12] Luo D., Yang J., Krstajic M., Ribarsky W., Keim D.: Eventriver: Visually exploring text collections with temporal references. *IEEE Trans. Vis. Comput. Graph. 18*, 1 (2012), 93–105. 3

[MCZZ08] Mei Q., Cai D., Zhang D., Zhai C.: Topic modeling with network regularization. In *Proceedings of the 17th International Conference on World Wide Web* (New York, NY, USA, 2008), WWW '08, ACM, pp. 101–110. URL: http://doi.acm.org/10.1145/1367497.1367512, doi:10.1145/1367497.1367512. 1

[Mon17] Montani I.: An open-source named entity visualiser for the modern web, 2017. https://explosion.ai/blog/displacy-ent-named-entity-visualizer [Accessed date: April 10, 2017]. 3

[MSM12] Meyer M., Sedlmair M., Munzner T.: The four-level nested model revisited: Blocks and guidelines. In *Proceedings of the 2012 BELIV Workshop: Beyond Time and Errors - Novel Evaluation Methods for Visualization* (New York, NY, USA, 2012), BELIV '12, ACM, pp. 11:1–11:6. URL: http://doi.acm.org/10.1145/2442576.2442587, doi:10.1145/2442576.2442587. 2

[Mun09] Munzner T.: A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics 15*, 6 (Nov. 2009), 921–928. URL: http://dx.doi.org/10.1109/TVCG.2009.111, doi:10.1109/TVCG.2009.111. 2

[New06] Newman M. E.: Modularity and community structure in networks. *Proceedings of the national academy of sciences 103*, 23 (2006), 8577–8582. 1

[PBMW99] Page L., Brin S., Motwani R., Winograd T.: *The PageRank Citation Ranking: Bringing Order to the Web.* Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120. URL: http://ilpubs.stanford.edu:8090/422/. 2

[RN11] Rotta R., Noack A.: Multilevel local search algorithms for modularity clustering. *J. Exp. Algorithmics 16* (July 2011), 2.3:2.1–2.3:2.27. URL: http://doi.acm.org/10.1145/1963190.1970376, doi:10.1145/1963190.1970376. 1

[WLS*10] Wei F., Liu S., Song Y., Pan S., Zhou M. X., Qian W., Shi L., Tan L., Zhang Q.: Tiara: a visual exploratory text analytic system. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (2010), ACM, pp. 153–162. 1

[WM06] Wang X., McCallum A.: Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (2006), ACM, pp. 424–433. 1

[ZLZG13] Zhang Z., Li Q., Zeng D., Gao H.: User community discovery from multi-relational networks. *Decision Support Systems 54*, 2 (2013), 870–879. 1