ETH zürich

ETH AI Center

Interactive Visual Analytics on TED Talks

Yaqi Qin, Tianyi Liu, Tianyang Xu, Aojie Yu

Department of Computer Science, ETH Zurich, CH

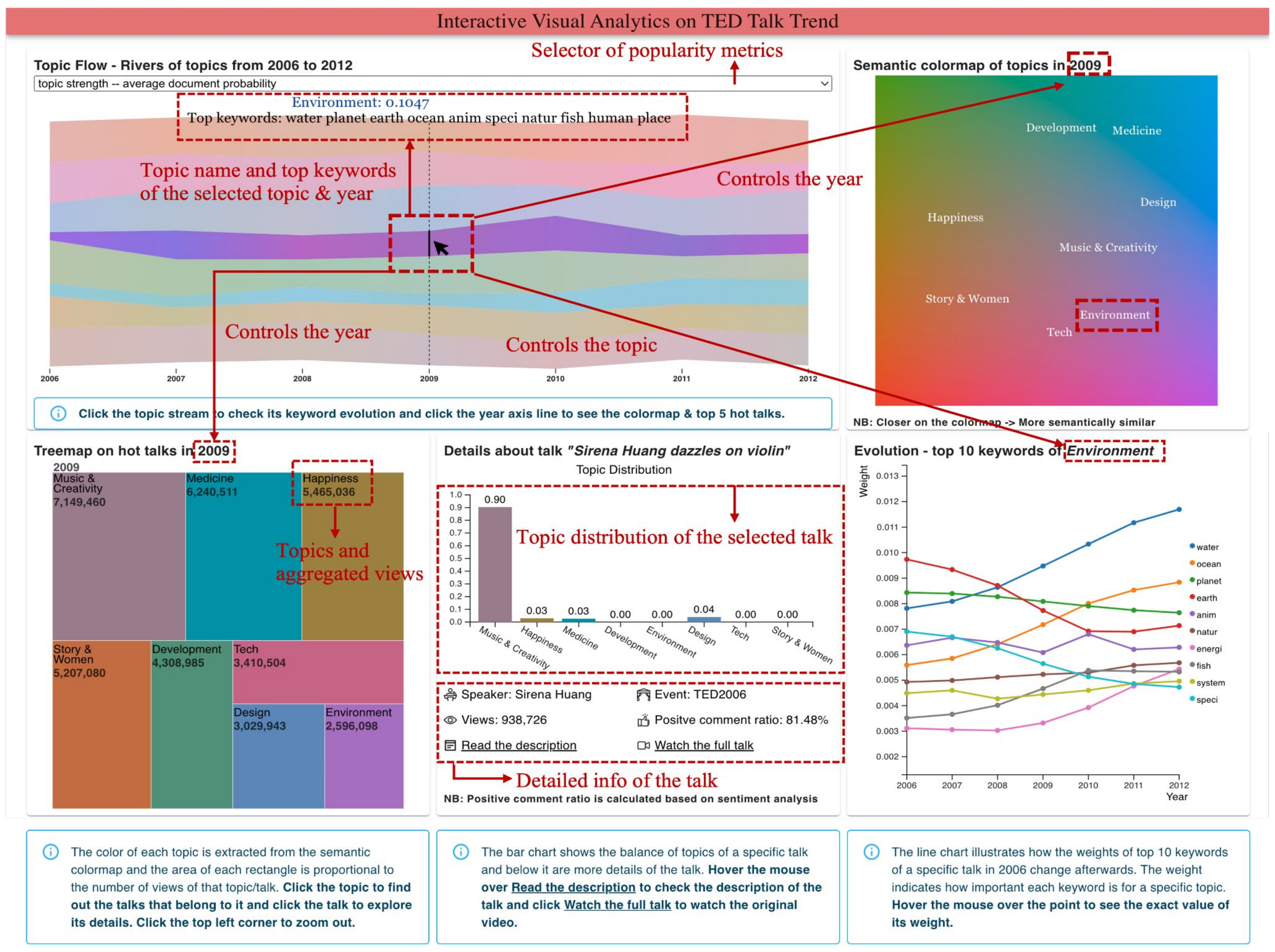


Figure 1. TED talk trend dashboard overview with 5 components

1 Problem Characterization

1.1 User

- TED foundation & event holders
  - Keep track of hot topics over time
  - Discover possible factors for a popular talk to attract audiences
- Social science researchers
  - Analyze public attention and topic competitiveness

1.2 Data

- TED talk dataset of talks published between 2006 and 2012 [3]
  - 1149 talks from 960 speakers and 69,023 registered users that have made about 200,000 comments
  - Data fields: identifier, title, transcript, publication date, number of views, related tags, user comments threads, etc.

1.3 Task

- Identify hot topics and study how they evolve over time → Topic flow component
- Identify hot talks for selected topics → Hot talks component, Talk details component
- Identify most important keywords and study how they evolve over time → Keyword evolution component
- Analyze topic correlation → Colormap component

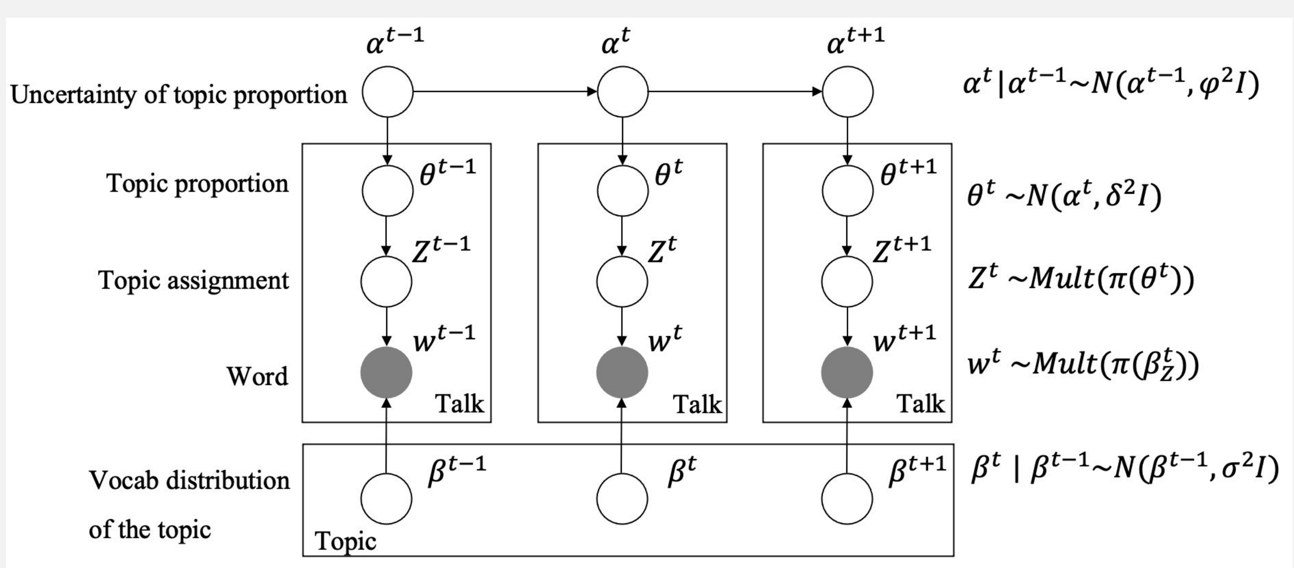
2 Topic Modelling

2.1 Data preprocessing for transcripts of talks

- Removing duplicates → Data cleaning → Extracting POS [NN, ADJ, VERB] → Removing stopwords → Stemming → Removing frequent words → Vocabulary of over 30,000 words

2.2 Dynamic LDA

- Use a dynamic Latent Dirichlet Allocation (LDA) topic model which captures the evolution of topics in a sequentially organized corpus of talks [1]
- Talks are grouped by year, and each year's talks arise from a set of topics that evolved from last year's topics
- Data model:
  i. topic[t]: a probability distribution over the vocabulary at time step t
  ii. talk: a probability distribution over all topics



2.3 Quantitative evaluation of topic popularity

Metric Name	Equation	Interpretation
Topic Strength (ST)	$ST_k^t = \frac{1}{\lvert D^t\rvert}\sum_{d\in D^t} P(k\mid d)$	Average topic probability of all talks
Topic coverage (CO)	$\mu_k^t = \sum_{d\in D^t} P(k\mid d)\times len(d) \Big/ \sum_{d\in D^t} len(d)$ $\sigma_k^t = \sqrt{\sum_{d\in D^t} (P(k\mid d)-\mu_k^t)^2\times len(d) \Big/ \sum_{d\in D^t} len(d)}$ $CO_k^t = [\mu_k^t]^{l1}\times[\sigma_k^t]^{l2}$	Measures topic on its average content coverage and its coverage variance [2]
Positive ratio (PR)	$PR_k^t = \frac{1}{\lvert D^t\rvert}\sum_{d\in D^t} P(k\mid d)\times\frac{\lvert\{C[SENT=positive]\mid d\}\rvert}{\lvert\{C\mid d\}\rvert}$	Average ratio of positive comments of talks under certain topic

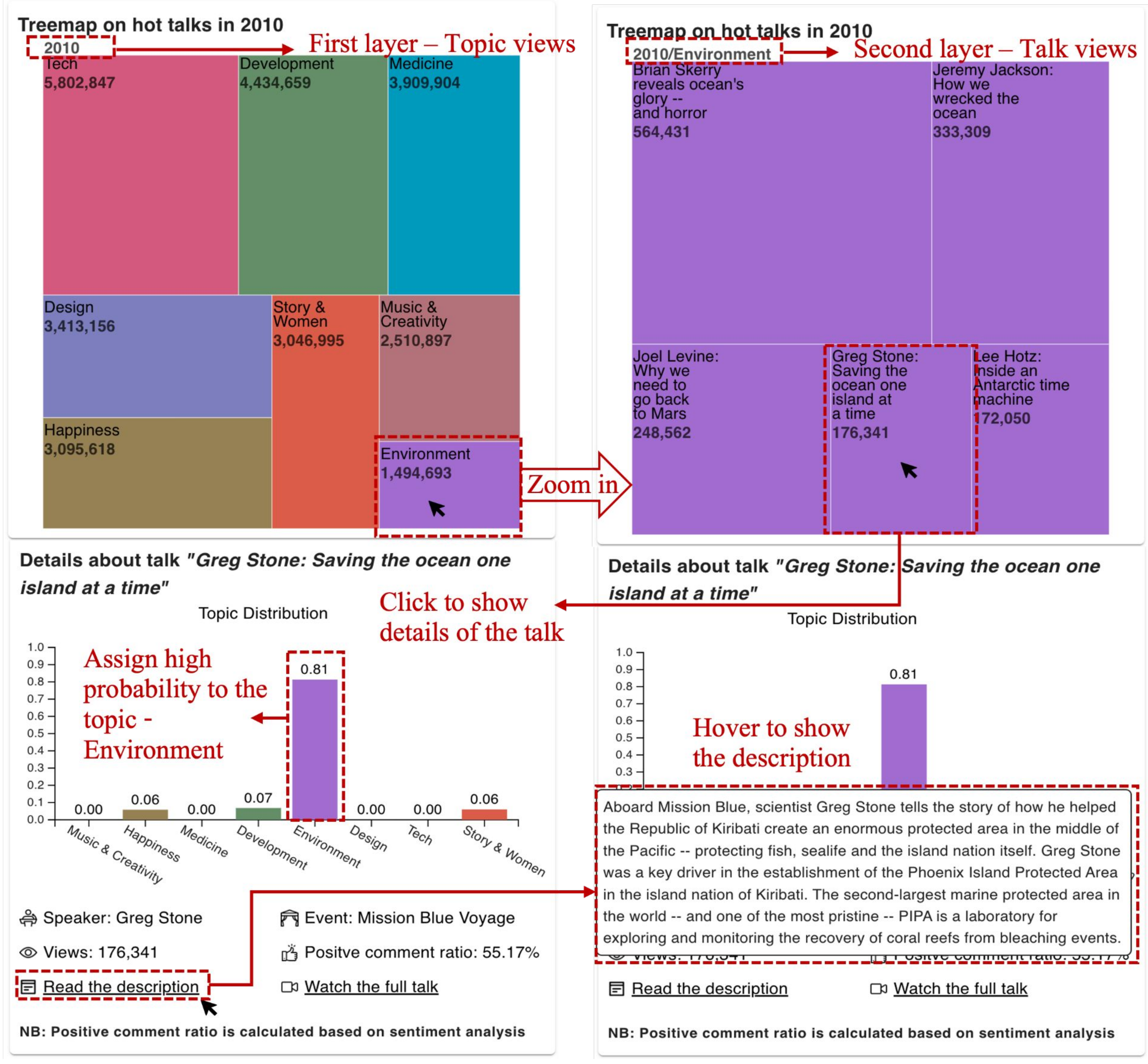


Figure 2. An example of interaction: Two layers of the hot talks component and its interaction with the talk details component

3 Sentiment Analysis

3.1 Data preprocessing for comments of talks

- Including only first level comments, i.e. excluding all the replies to other comments, because the target of their polarity judgment is uncertain (comments on comments rather than on the talk) [4]

3.2 Sentiment prediction

- Use Valence Aware Dictionary for sEntiment Reasoning (VADER) to calculate the sentiment score of each comment

3.3 Metrics

- Use the compound score (between 1 and -1) computed from the positive, negative and neutral score to determine the sentiment of each comment
- Rule: compound score > 0.4 → positive, compound score < -0.4 → negative, otherwise → neutral
- Count the number of positive, negative and neutral comments of each talk

4 Visualization

4.1 Topic flow component

- Use **stream graph** to display the evolution of topics based on a certain popularity metric
- Hover to highlight a specific stream, with topic name and keywords
- Click the stream or the vertical year axis line to change the colormap, the treemap, the bar chart, and the line chart

4.2 Colormap component

- Use distribution over vocabulary as the embedding vector of each topic of the selected year
- Use t-Distributed Stochastic Neighbor Embedding (t-SNE) to reduce the dimension to 2D, and project the 2D embedding of each topic onto a colormap to encode semantic similarity as color gradients
- Color the stream graph, the treemap and the bar chart

4.3 Hot talks component

- Use **zoomable treemap** to visualize the year-topic-hot video hierarchy of the selected year in stream graph, with the area of each block proportional to the aggregated number of views of each topic or talk
- Click the topic block to zoom in, click the top left to zoom out to the topics, and click the talk block to display the detailed view of the talk

4.4 Talk details component

- Use **bar chart** to display the topic distribution of the selected talk
- Provide other details including event, positive comment ratio, the description of the talk and the link to the original video

4.5 Keyword evolution component

- Use **line chart** to display the change in the importance of the top 10 keywords associated with the selected topic in 2006
- Hover to show the exact weight of a specific keyword

5 References

[1] David M. Blei and John D. Lafferty. "Dynamic topic models", Proceedings of the 23rd international conference on Machine learning (ICML '06), 113–120, 2006

[2] Liu Shixia, et al. "Tiara: Interactive, topic-based visual text summarization and analysis", ACM Transactions on Intelligent Systems and Technology (TIST) 3.2 (2012): 1-28.

[3] Nikolaos Pappas and Andrei Popescu-BelisPappas. "Combining Content with User Preferences for TED Lecture Recommendation", 11th International Workshop on Content Based Multimedia Indexing, IEEE, 2013

[4] Nikolaos Pappas and Andrei Popescu-BelisPappas. "Sentiment Analysis of User Comments for One-Class Collaborative Filtering over TED Talks", 36th ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2013