## How Hierarchical Topics Evolve in Large Text Corpora

A visualization for how topics of texts change over time

<table>
<thead>
<tr>
<th>Case study</th>
<th>Edward Snowden and the PRISM scandal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idea</td>
<td>Make trees out of topics</td>
</tr>
<tr>
<td>Problem</td>
<td>Dense graphs are tough to read and navigate through</td>
</tr>
<tr>
<td>Solution</td>
<td>Facet, details on demand, and align</td>
</tr>
<tr>
<td>Criticism</td>
<td></td>
</tr>
</tbody>
</table>

Weiwei Cui, Member, IEEE, Shixia Liu, Senior Member, IEEE, Zhuofeng Wu, Hao Wei
Case study

PRISM scandal: Edward Snowden leaked documents

Example topics

• “Snowden” vs “NSA”
• “Traitor” vs “Hero”

Too broad?
XKCD Inspiration for a similar system, this one uses same ideas.
Idea

**Topic trees as they evolve**

![Diagram showing topic trees evolving over time.]

**Problems**

- Topics are not at same level
- Changes are tough to track
- Users get lost when drilling down
Solution

Overview

Workflow
- Visualize
- Analyze
- Refine

Using
- Algorithm
- Domain knowledge
- Interaction

Iterative analysis
Tree cut

Every path from the root of the tree to a leaf will contain exactly one node from the cut.
Align twice
- For a unit of time
- For a level of the tree
Word cloud exposes structure of visualization
Solution  **Cut and repeat**

Break large topic into smaller topics
  - Large abstract topics may not be meaningful
  - Algorithm may not choose correctly

Iterate
  - More in line with how people actually think
So, how do you glean meaning from this?

(a) a new topic is emerging
(b) a topic is still active but changes slowly
(c) a topic is active, but changes immensely
(d) a momentary topic emerges and disappears rapidly.
What does this tell us about the news cycle?
What part is most important?
Which story is most important?
What aren’t we seeing?
Analysis  Criticism

**Good**
- Lowers cognitive load
- Manual manipulation makes sense
- Supports natural exploration process

**OK, maybe not a problem**
- Not really an algorithmic solution
- Requires domain knowledge to use

**Bad**
- Screen real estate ≠ importance
- Absolute Y-Pos means nothing, but it looks like it should
- Crossing, do we have to accept bad semantics?