Threads: Trees and Clusters

Scott Hendrickson
Principal Data Scientist, Gnip
@DrSkippy27

October 3, 2013

What do we talk about when they talk about X?

Apologies: Raymond Carver

Disqus Tree Structures

articles ← comments

comments ← comments

Disqus Threads

- 7 weeks
- Key words: “texting,” “driving” and variants
- Select top threads based on mentions
- 61,406 comments from 365 threads

Disqus Topic Model Approach

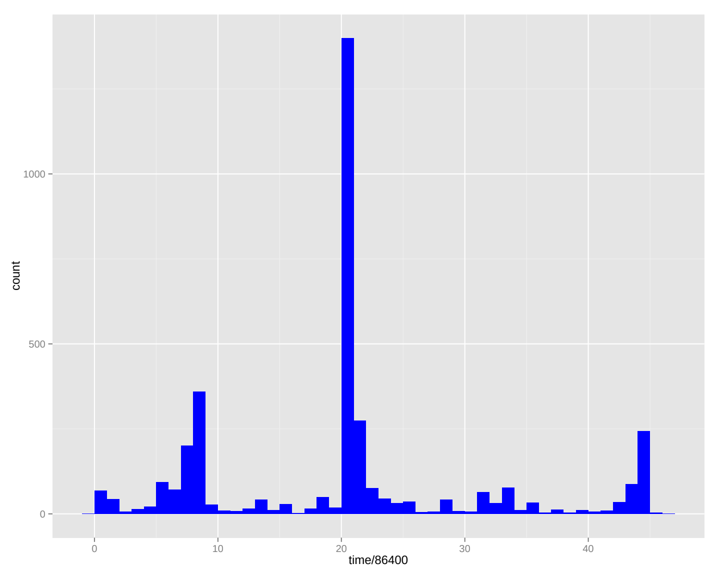
- Find comments that mention key words
- Corpus of comments (across many threads)
- tf-idf matrix: terms × comments
- LSI (rotate space to align with “important” dimensions, cut dimensions)
- K-means (quick-and-dirty clustering in reduced dimensional space)
- …rinse and repeat (looking for distinction and cohesion)

Disqus Topic Model

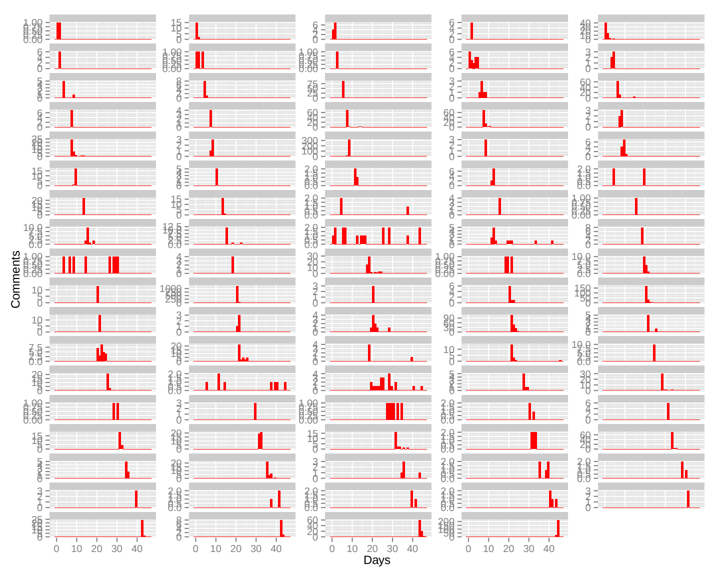
- Same 7 weeks; same keywords
- 32,856 comments from 16,886 threads
- LSI: 500 features → 80 features
- K-means: 80 clusters as topics (?!)

Focus on the intersection of Thread and Topic models

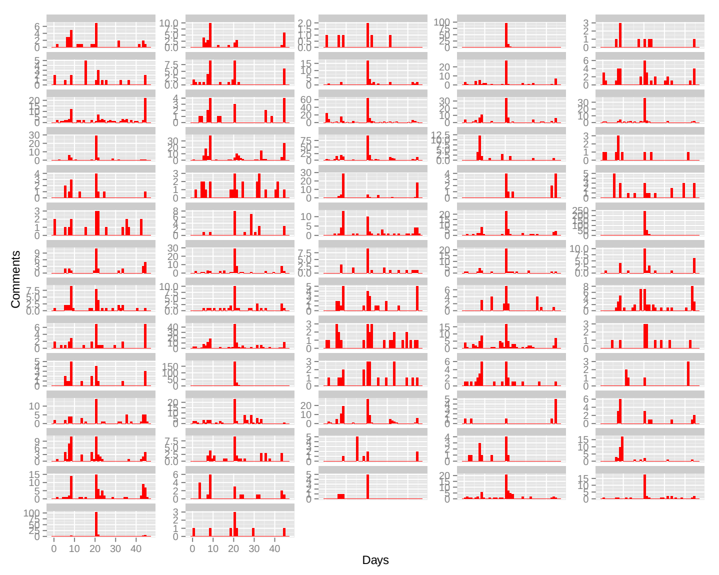
Comments with key words over time



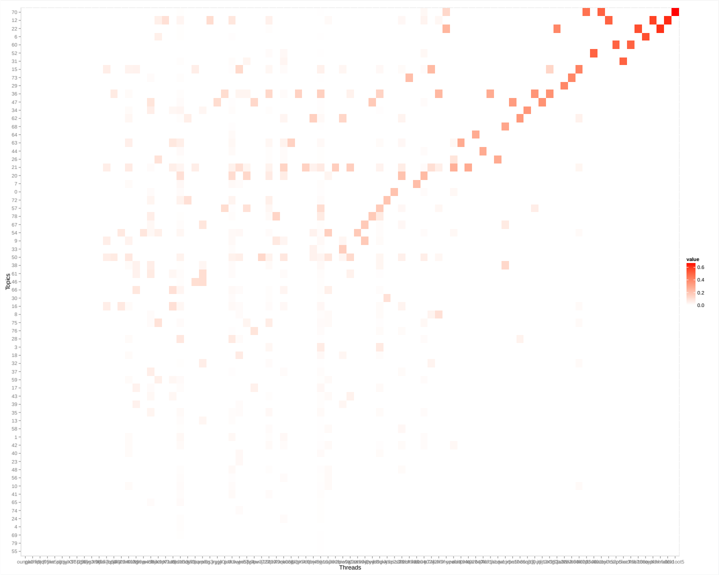
Disqus Thread Activity over Time



Disqus Topics Activity over Time



Dominant Topics × Threads?



When we talk about texting and driving, we talk about …

- Topic 12: poor graphic design
- Topic 50: fake ids and fake drivers licenses
- Topic 58: health/accident insurance
- Topic 62: drunk drivers
- Topic 64: buses and bus drivers
- Topic 67: bikes, bike lanes
- Topic 68: trucks and truck drivers

Thank you!

- Presentation, data, vis. code at:
  http://github.com/DrSkippy27/Disqus-Lightening-2013