Data Science Case Study:

Visualizing Reddit Data

Justin Skycak

January 13, 2017

Goal

Generate a visualization of Reddit population segments that is as granular as possible, while still being interpretable.

Complexity

Reddit consists of over 200 million users commenting in nearly a million subreddits.

Tradeoffs with Conventional Approaches

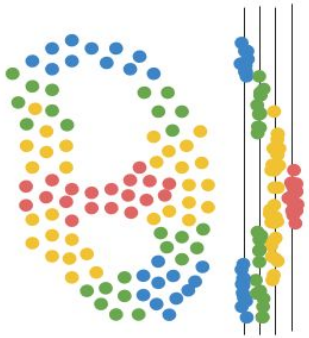
- **Network plotting**
  - Very granular
  - Hard to interpret
- **Statistical models (e.g. GMM)**
  - Easy to interpret
  - Computationally expensive (parameter blowup)
- **Pointwise plotting (e.g. PCA)**
  - Computationally inexpensive
  - Reduces granularity

Plan

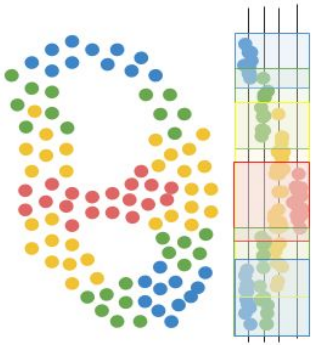
- **Visualization = TDA network**
  - Uses clusters (rather than individuals) as nodes → interpretable, yet granular
  - O(n^2) → not exponentially expensive

Mapper Refresher

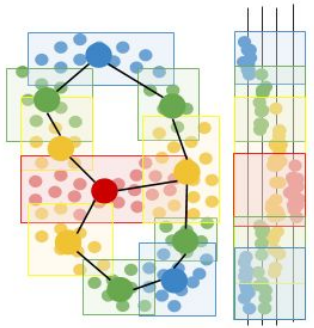
1. Stretch out the data along one dimension



2. Chop it into pieces



3. Cluster within each piece



Two Inputs to Mapper

1. **What will we use as primary records?**
2. **How will we compute the similarity between these records?**

Ideal Approach

1. **Primary records = Reddit users**
2. **Similarity = shared comment percentage**
    - User A: 60 comments to r/programming, 40 comments to r/AskReddit → [60%, 40%]
    - User B: 10 comments to r/programming, 30 comments to r/AskReddit → [25%, 75%]
    - Similarity(A,B) = shared percentage = 25% + 40% = 65%

Getting the Comments Data

- First tried Reddit API
  - Too slow (<1000 users/hour)
- Then queried public Reddit dataset on BigQuery
  - Success!

Computing Similarities

- N users → N^2 similarity measurements
  - Limited to N = 50,000 users for computation to finish in reasonable time
  - Not sure if sufficient, but couldn't think of a quick workaround

Computation Issues with TDA

- **To complete in reasonable time, R implementation of TDA can take at most a ~10,000 x 10,000 distance matrix**
  - **Is 10,000 users enough? Can't think of a quick workaround, so let's cross our fingers and continue.**
- **How to test if 10,000 sample is sufficiently large?**
  - **Is the resulting topological network stable? I.e. if you choose a different sample of 10,000 users, will the network be similar?**
- **Unfortunately, 10,000 sample is too small.**

Second-Best Approach

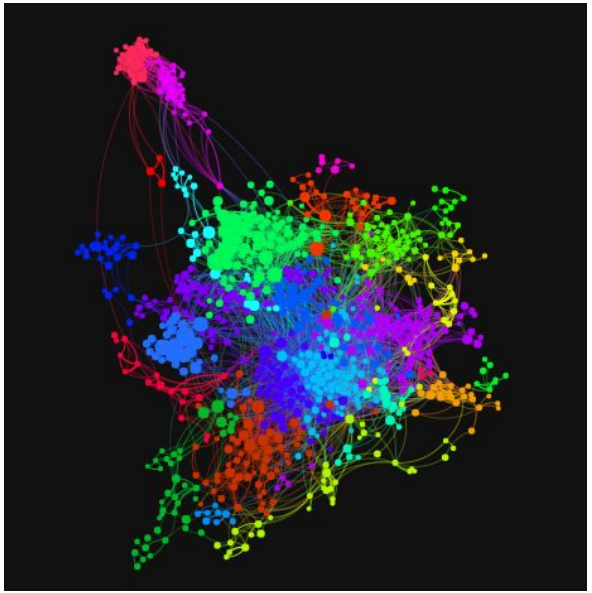
1. **Primary records = 10,000 most popular subreddits**
2. **Similarity = percentage of shared commenters**
    - Subreddit A: 1,000,000 commenters total
    - Subreddit B: 500,000 commenters total
    - Number of shared commenters: 300,000

    - Percentage of shared commenters
        = (number of shared commenters) / (number of unique commenters)
        = 300,000 / (1,000,000 + 500,000 - 300,000)
        = 300,000 / 1,200,000
        = 0.25

Data-Gathering Shortcut

- **People have done this analysis before to create Reddit networks - they just haven't tried TDA before**

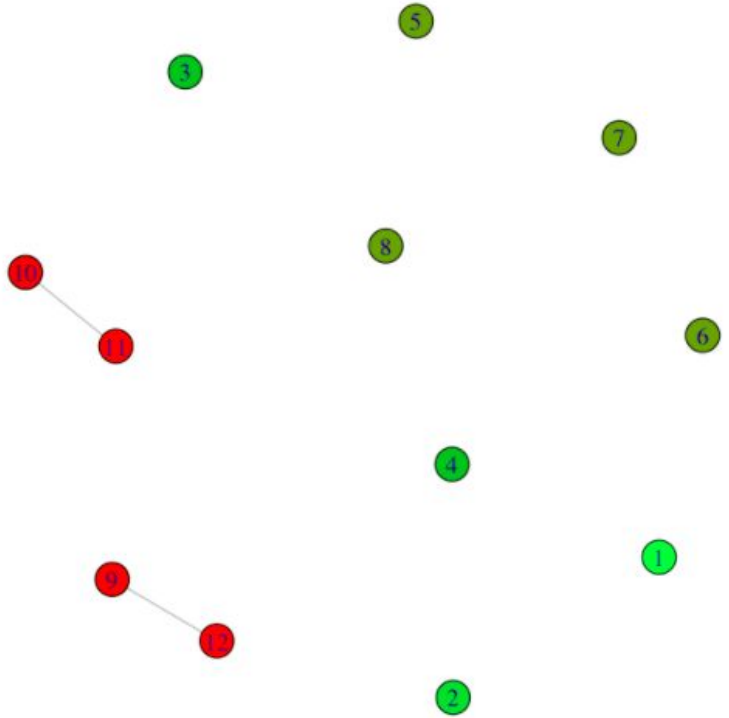
- **Shortcut: Download their graph xml file and scrape the subreddit names and edge weights**



http://www.jacobsilterra.com/subreddit_map/network/index.html#

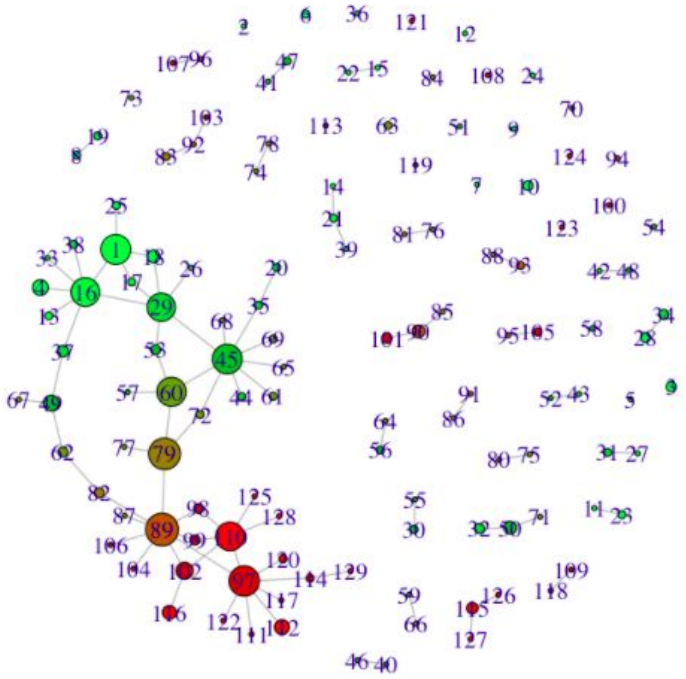
Initial Result

- **Overall boring network, but some interesting groupings:**
  - **1:** guns, firearms
  - **4:** cars, autos
  - **8:** windows, microsoft
  - **9/12:** ukpolitics, unitedkingdom, scotland
  - **Many other nodes:** different varieties of porn



Result after Distance Transformation

- **Full disclosure: this was a lucky accident**
- **More interesting groupings:**
  - **112:** buildapcsales, mechanicalkeyboards, shouldibuythisgame, mountandblade, cynicalbrit, gamedeals, battlestations, paradoxplaza
  - **98:** design, wordpress, web_design, startups
  - **44:** bmw, cars, roadcam
  - **38:** aw, nyc, cityporn (not actually porn, but beautiful pictures of cities)
  - **18:** baseball, nfl, eagles, mma, chibears, cfb
  - **116/4/9:** more porn



Why did Distance Transformation affect Results?

- **In theory, continuous transformations shouldn't have any effect on topology**
- **But in practice, having a finite number of records can invalidate this conclusion**
  - Take the interval [0,1] and raise each number to the 100th power. You get the same interval, [0,1].
  - Take a sample {0, .1, …, .9, 1} and do the same thing. Now you get {0, 0, …, .00003, 1}.
    - When you use Mapper on [0, .1, …, .9, 1], you get a bunch of connected clusters:
      (0,.1,.2)-----(.2,.3,.4)-----(.4,.5,.6)-----(and so on…)
    - When you use Mapper on {0, 0, …, .00003, 1} you get two disconnected clusters:
      (0, 0, …, .00003) (1)

Learning Points

- **TDA claims not to depend on the choice distance metrics, but this is only true in dense spaces.**
  Real-world data is discrete and does not satisfy this assumption. Therefore, the choice of distance metric matters.

- **Choosing distance metrics is more an art than a science.**
  The best choice depends on what you're looking for, and in exploratory data science you often don't know what you're looking for until you've found it.

- **I learned some basic data engineering in the realm of APIs, SQL, and BigQuery.**

Questions? :)