Thinking Critically about Digital Data Collection

Twitter and Beyond

Rebekah Tromble
Leiden University

POLICE
Stop Charging
or We Use Force

Motivation

- Understanding what we actually get
  - Corporate data providers
  - Third-party intermediaries
  - Biases generated
- Critically engaging the ethics of data collection and management

Twitter Basics

- Why Twitter?
  - After all…
    - Facebook & Instagram much more popular
    - Not representative (not public opinion)
  - But…
    - 96-98% public
    - Opinion leaders
    - Media coverage

Twitter Basics –
Means of Collecting the Data

- Application Programming Interfaces (APIs)
  - Firehose – real-time, 100%, cost-prohibitive
  - Streaming – real-time, sample
  - Search/Rest – historical, but with significant limitations
- Archive – All tweets since June 2006…sort of…
- Scraping

Twitter Basics –
Means of Collecting the Data

- Third-Party Services
  - Firehose
    - CrimsonHexagon – Full Firehose, limited access to content
    - DiscoverText – PowerTrack
  - Streaming – TCAT
  - Search/Rest – DiscoverText, Node XL, NCapture for Nvivo, TCAT
  - Archive – Gnip, Sifter (DiscoverText)

Understanding the APIs

- Streaming (Keyword queries)
  - Real time capture:
  - Can capture up to 1% of global volume – rate limits
    - Issue/event is popular
    - Americans go to sleep/on vacation

Understanding the APIs

- Search/Rest (Keyword queries)
  - Historical capture by keyword or username
  - Significant limitations:
    - Up to 18,000 tweets over the last ~7-10-day period, whichever limit is reached first.
    - Up to 180 calls every 15 minutes.
    - Captures far less than 100% (“top” tweets).

Understanding the APIs

- Firehose – real-time, 100%
  - Only accessible through official Twitter partner.
  - Cost-prohibitive.
  - Designed for corporate use.
  - Some services won’t let you see the tweets.
- Archive – not actually an API
  - Not truly a record of all tweets.
  - Terms of service require everyone to remove deleted tweets.
    - Special arrangement with PolitWoops

Data Collection

- U.S. election
  - 48 hours: 8-9 November 2016
  - Keyword query: govgaryjohnson OR drjillstein OR evan_mcmullin
  - PowerTrack – real time, queried from the firehose
    - 226,118 tweets
  - Streaming API – real time, no rate limits hit
    - 185,490 tweets = 82.0%
  - Search API – maximum calls, 8-17 November
    - 112,758 tweets = 49.9%

Research Question 1

- What bias is introduced using different APIs?
  - Extracted @mentions and usernames
  - Compared “top” lists using Kendall’s Tau

	Mentions		Usernames	
Top #	PowerTrack - Stream	PowerTrack - Search	PowerTrack - Stream	PowerTrack - Search
10	0.7778	0.2444	0.7333	0.6000
25	0.8467	0.4667	0.86	0.5776
50	0.8237	0.6131	0.882	0.6032
100	0.8179	0.5823	0.9008	0.5702
250	0.8152	0.5557	0.8528	0.5262
500	0.8119	0.5145	0.8577	0.5282
1000	0.8004	0.5249	0.835	0.5376

Research Question 2

- What factors drive API samples?
- Logit regression
  - User characteristic variables
    - How prolific? (status count)
    - How popular? (follower count)
    - How engaged? (friend count)
  - Tweet characteristic variables
    - Originality? (retweet)
    - Engagement w/ others? (mentions count)
    - Engagement in discourse? (hashtag count)
    - Content richness? (multimedia)

Analysis

- Ran 40 models
  - Step-wise test of interaction effects
  - Simplest proved best.

	Search		Streaming	
Variable	Coeff	Odds Ratio	Coeff	Odds Ratio
Status count	9.37E-07***	1.0000009	6.05E-07***	1.0000006
Followers	-9.80E-08***	0.9999999	1.24E-08	1.0000000
Friends	6.28E-06***	1.0000063	3.77E-07	1.0000004
Retweet	-3.09E-01***	0.7344833	-5.67E-01***	0.5672836
Mention count	-4.08E-02***	0.9599965	-3.63E-02***	0.9643693
Hashtag count	1.24E-01***	1.1323	3.08E-02***	1.0312944
Multimedia	-5.97E-03	0.9940459	4.58E-01***	1.5816395
Intercept	-2.08E-01***	0.8118603	1.83E+00***	6.2159437

	Search		Streaming	
Variable	Coeff	Odds Ratio	Coeff	Odds Ratio
Status count	9.37E-07***	1.0000009	6.05E-07***	1.0000006
Followers	-9.80E-08***	0.9999999	1.24E-08	1.0000000
Friends	6.28E-06***	1.0000063	3.77E-07	1.0000004
Retweet	-3.09E-01***	0.7344833	-5.67E-01***	0.5672836
Mention count	-4.08E-02***	0.9599965	-3.63E-02***	0.9643693
Hashtag count	1.24E-01***	1.1323	3.08E-02***	1.0312944
Multimedia	-5.97E-03	0.9940459	4.58E-01***	1.5816395
Intercept	-2.08E-01***	0.8118603	1.83E+00***	6.2159437

	Search		Streaming	
Variable	Coeff	Odds Ratio	Coeff	Odds Ratio
Status count	9.37E-07***	1.0000009	6.05E-07***	1.0000006
Followers	-9.80E-08***	0.9999999	1.24E-08	1.0000000
Friends	6.28E-06***	1.0000063	3.77E-07	1.0000004
Retweet	-3.09E-01***	0.7344833	-5.67E-01***	0.5672836
Mention count	-4.08E-02***	0.9599965	-3.63E-02***	0.9643693
Hashtag count	1.24E-01***	1.1323	3.08E-02***	1.0312944
Multimedia	-5.97E-03	0.9940459	4.58E-01***	1.5816395
Intercept	-2.08E-01***	0.8118603	1.83E+00***	6.2159437

(Tentative) Conclusions

- Content matters
- User does not
- We are looking at especially “rich” content. This has clear consequences for interpretation.

Research Question 3

- How does digital data decay over time?
- What are some of the ethical implications of digital data collection?

Beyond Twitter

- Timing of data collection always matters
- Facebook
  - Far more private content
  - Can scrape Facebook groups and pages – raises serious ethical concerns
- Reddit
- SnapChat
- WayBackMachine
  - No clue about the algorithm

Data Collection Demonstration

- Advanced Search + Scraping (+ Rest API)
  - Twitter Advanced Search:
    - https://twitter.com/search-advanced
    - Key Tips:
      - Search day before and day after
      - Go to “Latest” results, not “Top”
  - Web Scraper tutorials:
    - http://webscraper.io/tutorials
  - Rest API: showStatus, lookupstatuses

Thank you!