Collecting Social Media Data

Slides: go.gwu.edu/socialmediadata

Dan Kerchner
kerchner@gwu.edu

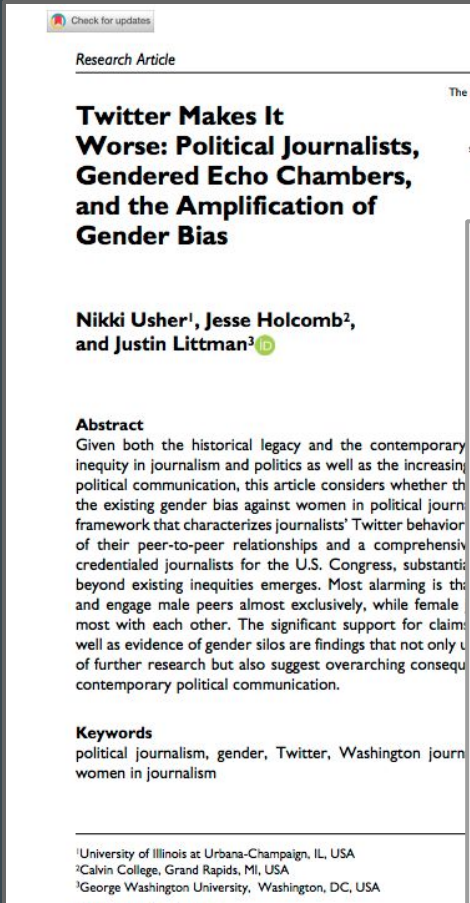
Laura Wrubel
lwrubel@gwu.edu

If you'd like to follow along for the optional hands-on portion, sign into the GW VPN.

Agenda

- Overview of social media APIs and data formats
- Twitter's API in depth
- Collecting new datasets
  - Hands-on: Social Feed Manager
- Using existing datasets
  - Demo: TweetSets
- Approaches for other social media platforms
- Ethics of social media collecting

Social media research



Check for updates

Research Article

The International Journal of Press/Politics
2018, Vol. 23(3) 324–344
© The Author(s) 2018
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1940161218781254
journals.sagepub.com/home/hij
SAGE

Twitter Makes It Worse: Political Journalists, Gendered Echo Chambers, and the Amplification of Gender Bias

Nikki Usher[1], Jesse Holcomb[2], and Justin Littman[3]

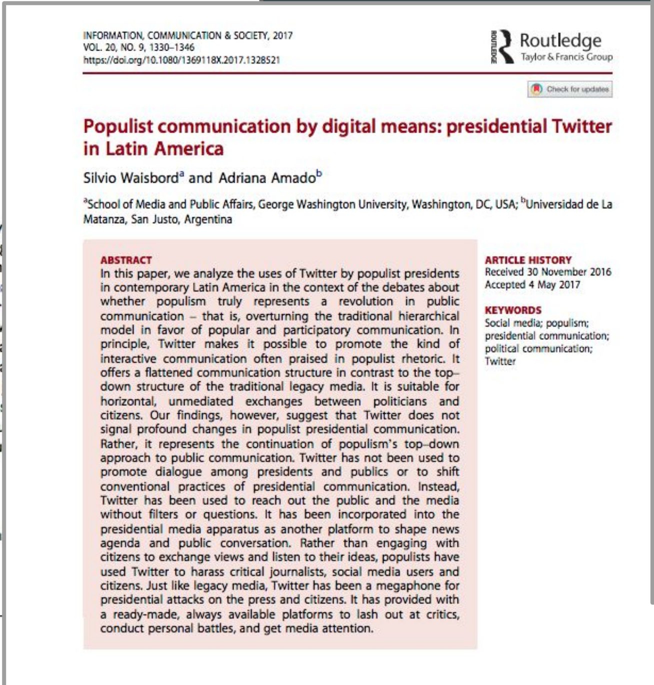
Abstract

Given both the historical legacy and the contemporary inequity in journalism and politics as well as the increasin political communication, this article considers whether th the existing gender bias against women in political journ framework that characterizes journalists' Twitter behavior of their peer-to-peer relationships and a comprehensiv credentialed journalists for the U.S. Congress, substanti beyond existing inequities emerges. Most alarming is th and engage male peers almost exclusively, while female most with each other. The significant support for claim well as evidence of gender silos are findings that not only of further research but also suggest overarching consequ contemporary political communication.

Keywords

political journalism, gender, Twitter, Washington journ women in journalism

[1] University of Illinois at Urbana-Champaign, IL, USA
[2] Calvin College, Grand Rapids, MI, USA
[3] George Washington University, Washington, DC, USA



INFORMATION, COMMUNICATION & SOCIETY, 2017
VOL. 20, NO. 9, 1330–1346
https://doi.org/10.1080/1369118X.2017.1328521
Routledge
Taylor & Francis Group

Check for updates

Populist communication by digital means: presidential Twitter in Latin America

Silvio Waisbord[a] and Adriana Amado[b]

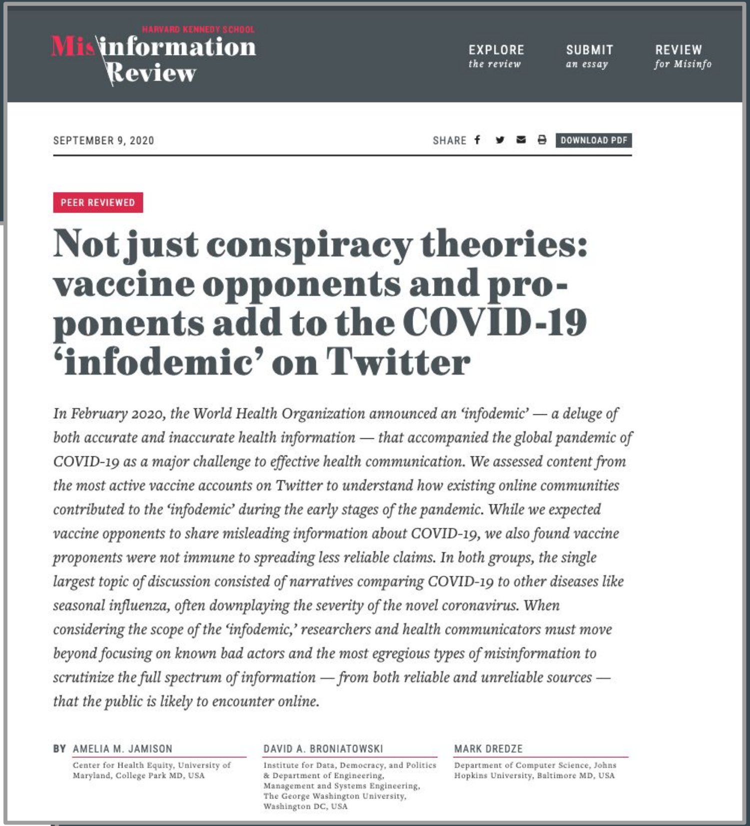
[a] School of Media and Public Affairs, George Washington University, Washington, DC, USA; [b] Universidad de La Matanza, San Justo, Argentina

ABSTRACT

In this paper, we analyze the uses of Twitter by populist presidents in contemporary Latin America in the context of the debates about whether populism truly represents a revolution in public communication – that is, overturning the traditional hierarchical model in favor of popular and participatory communication. In principle, Twitter makes it possible to promote the kind of interactive communication often praised in populist rhetoric. It offers a flattened communication structure in contrast to the top–down structure of the traditional legacy media. It is suitable for horizontal, unmediated exchanges between politicians and citizens. Our findings, however, suggest that Twitter does not signal profound changes in populist presidential communication. Rather, it represents the continuation of populism's top–down approach to public communication. Twitter has not been used to promote dialogue among presidents and publics or to shift conventional practices of presidential communication. Instead, Twitter has been used to reach out the public and the media without filters or questions. It has been incorporated into the presidential media apparatus as another platform to shape news agenda and public conversation. Rather than engaging with citizens to exchange views and listen to their ideas, populists have used Twitter to harass critical journalists, social media users and citizens. Just like legacy media, Twitter has been a megaphone for presidential attacks on the press and citizens. It has provided with a ready-made, always available platforms to lash out at critics, conduct personal battles, and get media attention.

ARTICLE HISTORY
Received 30 November 2016
Accepted 4 May 2017

KEYWORDS
Social media; populism; presidential communication; political communication; Twitter



HARVARD KENNEDY SCHOOL
Misinformation Review

SEPTEMBER 9, 2020

PEER REVIEWED

Not just conspiracy theories: vaccine opponents and pro-ponents add to the COVID-19 'infodemic' on Twitter

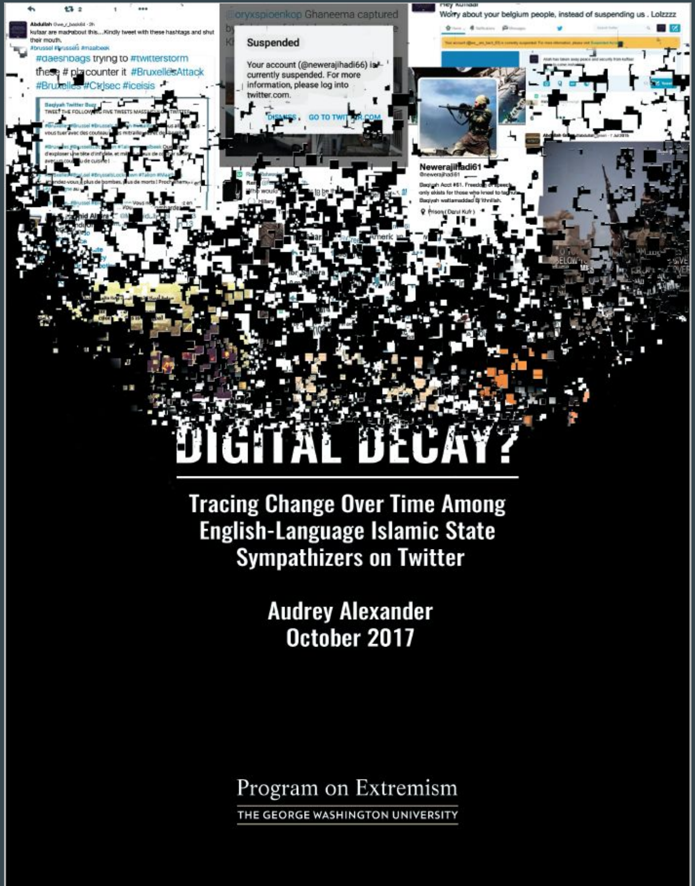
In February 2020, the World Health Organization announced an 'infodemic' — a deluge of both accurate and inaccurate health information — that accompanied the global pandemic of COVID-19 as a major challenge to effective health communication. We assessed content from the most active vaccine accounts on Twitter to understand how existing online communities contributed to the 'infodemic' during the early stages of the pandemic. While we expected vaccine opponents to share misleading information about COVID-19, we also found vaccine proponents were not immune to spreading less reliable claims. In both groups, the single largest topic of discussion consisted of narratives comparing COVID-19 to other diseases like seasonal influenza, often downplaying the severity of the novel coronavirus. When considering the scope of the 'infodemic,' researchers and health communicators must move beyond focusing on known bad actors and the most egregious types of misinformation to scrutinize the full spectrum of information — from both reliable and unreliable sources — that the public is likely to encounter online.

BY AMELIA M. JAMISON
Center for Health Equity, University of Maryland, College Park MD, USA

DAVID A. BRONIATOWSKI
Institute for Data, Democracy, and Politics & Department of Engineering, Management and Systems Engineering, The George Washington University, Washington DC, USA

MARK DREDZE
Department of Computer Science, Johns Hopkins University, Baltimore MD, USA

Social media research

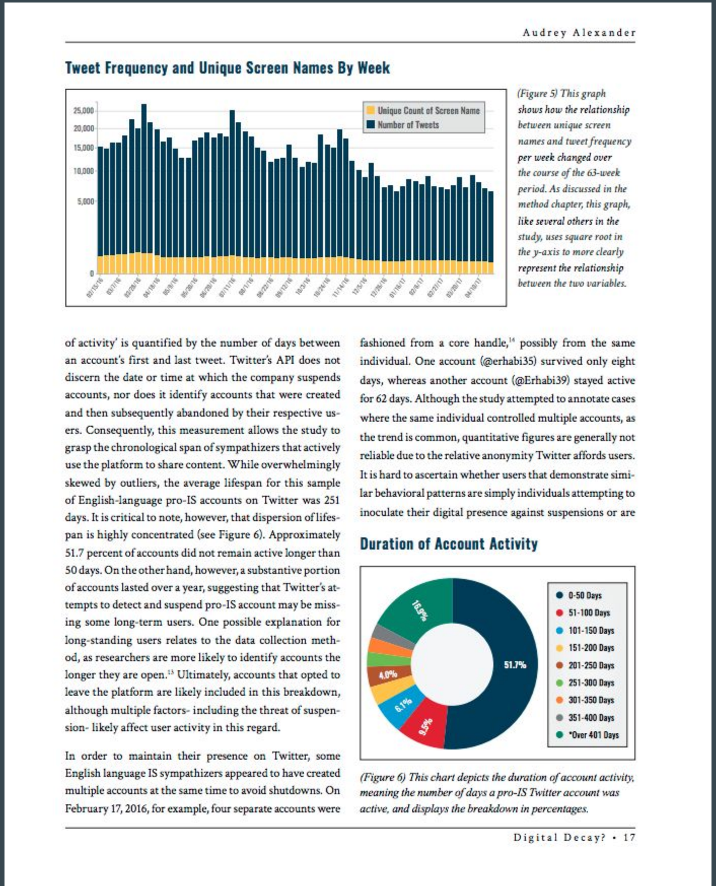


DIGITAL DECAY?

Tracing Change Over Time Among
English-Language Islamic State
Sympathizers on Twitter

Audrey Alexander
October 2017

Program on Extremism
THE GEORGE WASHINGTON UNIVERSITY



Tweet Frequency and Unique Screen Names By Week

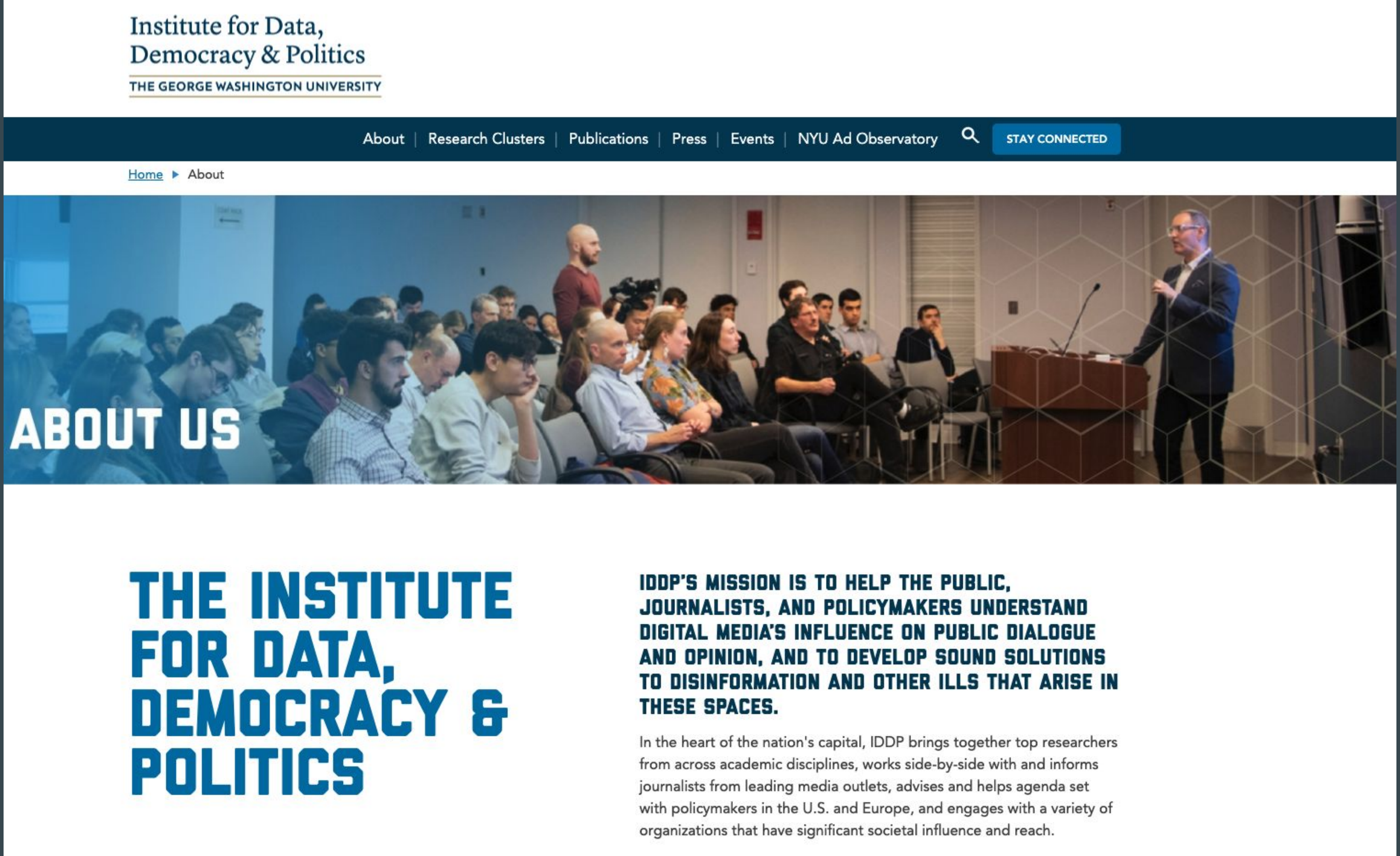
(Figure 5) This graph shows how the relationship between unique screen names and tweet frequency per week changed over the course of the 63-week period. As discussed in the method chapter, this graph, like several others in the study, uses square root in the y-axis to more clearly represent the relationship between the two variables.

of activity' is quantified by the number of days between an account's first and last tweet. Twitter's API does not discern the date or time at which the company suspends accounts, nor does it identify accounts that were created and then subsequently abandoned by their respective users. Consequently, this measurement allows the study to grasp the chronological span of sympathizers that actively use the platform to share content. While overwhelmingly skewed by outliers, the average lifespan for this sample of English-language pro-IS accounts on Twitter was 251 days. It is critical to note, however, that dispersion of lifespan is highly concentrated (see Figure 6). Approximately 51.7 percent of accounts did not remain active longer than 50 days. On the other hand, however, a substantive portion of accounts lasted over a year, suggesting that Twitter's attempts to detect and suspend pro-IS account may be missing some long-term users. One possible explanation for long-standing users relates to the data collection method, as researchers are more likely to identify accounts the longer they are open.[13] Ultimately, accounts that opted to leave the platform are likely included in this breakdown, although multiple factors- including the threat of suspension- likely affect user activity in this regard.

In order to maintain their presence on Twitter, some English language IS sympathizers appeared to have created multiple accounts at the same time to avoid shutdowns. On February 17, 2016, for example, four separate accounts were fashioned from a core handle,[14] possibly from the same individual. One account (@erhabi35) survived only eight days, whereas another account (@Erhabi39) stayed active for 62 days. Although the study attempted to annotate cases where the same individual controlled multiple accounts, as the trend is common, quantitative figures are generally not reliable due to the relative anonymity Twitter affords users. It is hard to ascertain whether users that demonstrate similar behavioral patterns are simply individuals attempting to inoculate their digital presence against suspensions or are

Duration of Account Activity

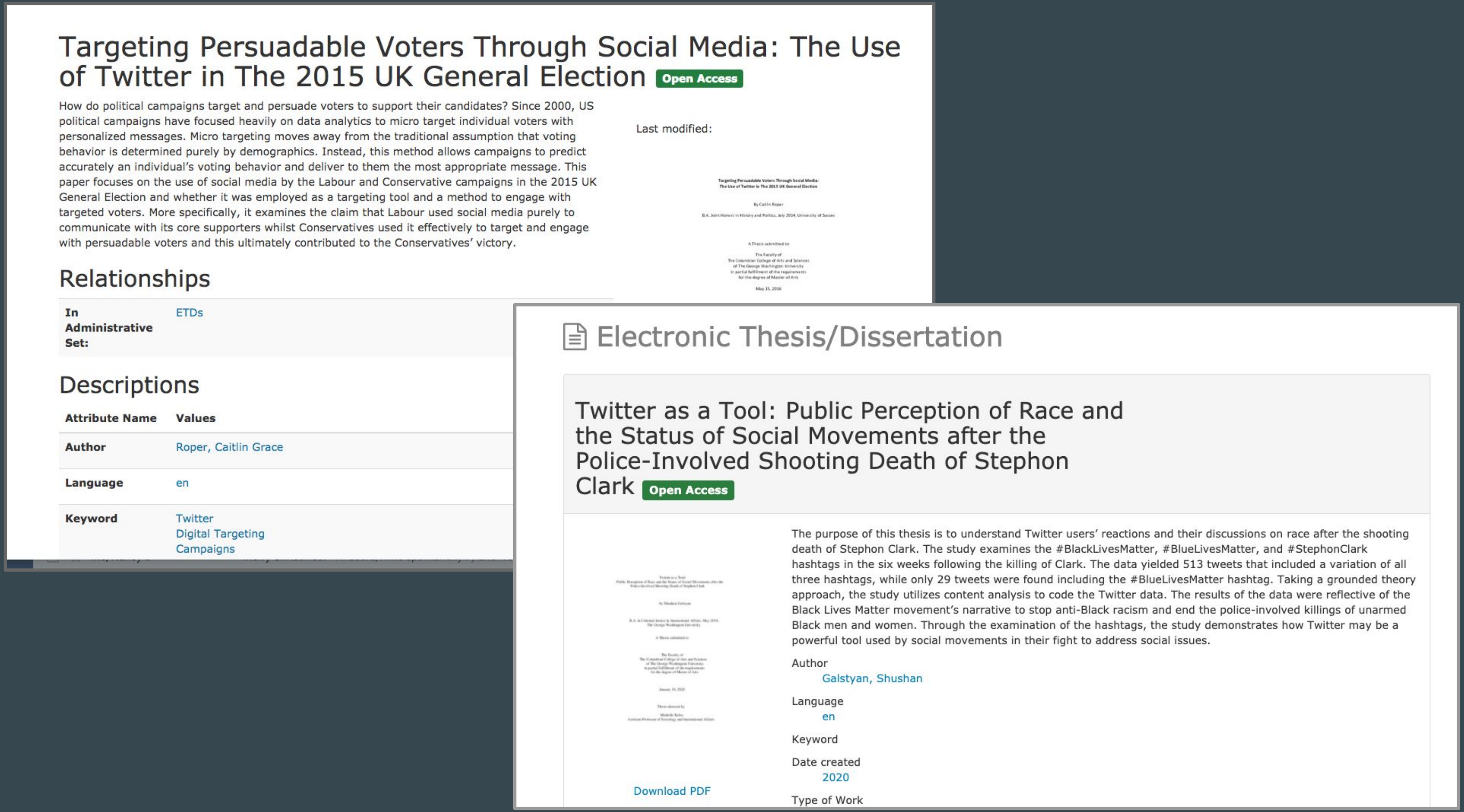
(Figure 6) This chart depicts the duration of account activity, meaning the number of days a pro-IS Twitter account was active, and displays the breakdown in percentages.



THE INSTITUTE FOR DATA, DEMOCRACY & POLITICS

IDDP'S MISSION IS TO HELP THE PUBLIC, JOURNALISTS, AND POLICYMAKERS UNDERSTAND DIGITAL MEDIA'S INFLUENCE ON PUBLIC DIALOGUE AND OPINION, AND TO DEVELOP SOUND SOLUTIONS TO DISINFORMATION AND OTHER ILLS THAT ARISE IN THESE SPACES.

In the heart of the nation's capital, IDDP brings together top researchers from across academic disciplines, works side-by-side with and informs journalists from leading media outlets, advises and helps agenda set with policymakers in the U.S. and Europe, and engages with a variety of organizations that have significant societal influence and reach.



Targeting Persuadable Voters Through Social Media: The Use of Twitter in The 2015 UK General Election

How do political campaigns target and persuade voters to support their candidates? Since 2000, US political campaigns have focused heavily on data analytics to micro target individual voters with personalized messages. Micro targeting moves away from the traditional assumption that voting behavior is determined purely by demographics. Instead, this method allows campaigns to predict accurately an individual's voting behavior and deliver to them the most appropriate message. This paper focuses on the use of social media by the Labour and Conservative campaigns in the 2015 UK General Election and whether it was employed as a targeting tool and a method to engage with targeted voters. More specifically, it examines the claim that Labour used social media purely to communicate with its core supporters whilst Conservatives used it effectively to target and engage with persuadable voters and this ultimately contributed to the Conservatives' victory.

Last modified:

Targeting Persuadable Voters Through Social Media:
The Use of Twitter in The 2015 UK General Election

By Caitlin Roper

B.A. Joint Honors in History and Politics, July 2014, University of Sussex

A Thesis submitted to

The Faculty of
The Columbian College of Arts and Sciences
of The George Washington University
in partial fulfillment of the requirements
for the degree of Master of Arts

May 15, 2016

Relationships

In Administrative Set:	ETDs

Descriptions

Attribute Name	Values
Author	Roper, Caitlin Grace
Language	en
Keyword	Twitter Digital Targeting Campaigns

Electronic Thesis/Dissertation

Twitter as a Tool: Public Perception of Race and the Status of Social Movements after the Police-Involved Shooting Death of Stephon Clark

Twitter as a Tool
Public Perception of Race and the Status of Social Movements after the
Police-Involved Shooting Death of Stephon Clark

by Shushan Galstyan

B.A. in Criminal Justice & International Affairs, May 2018
The George Washington University

A Thesis submitted to

The Faculty of
The Columbian College of Arts and Sciences
of The George Washington University
in partial fulfillment of the requirements
for the degree of Master of Arts

January 10, 2020

Thesis directed by

Michelle Kelso
Assistant Professor of Sociology and International Affairs

The purpose of this thesis is to understand Twitter users' reactions and their discussions on race after the shooting death of Stephon Clark. The study examines the #BlackLivesMatter, #BlueLivesMatter, and #StephonClark hashtags in the six weeks following the killing of Clark. The data yielded 513 tweets that included a variation of all three hashtags, while only 29 tweets were found including the #BlueLivesMatter hashtag. Taking a grounded theory approach, the study utilizes content analysis to code the Twitter data. The results of the data were reflective of the Black Lives Matter movement's narrative to stop anti-Black racism and end the police-involved killings of unarmed Black men and women. Through the examination of the hashtags, the study demonstrates how Twitter may be a powerful tool used by social movements in their fight to address social issues.

Author
Galstyan, Shushan

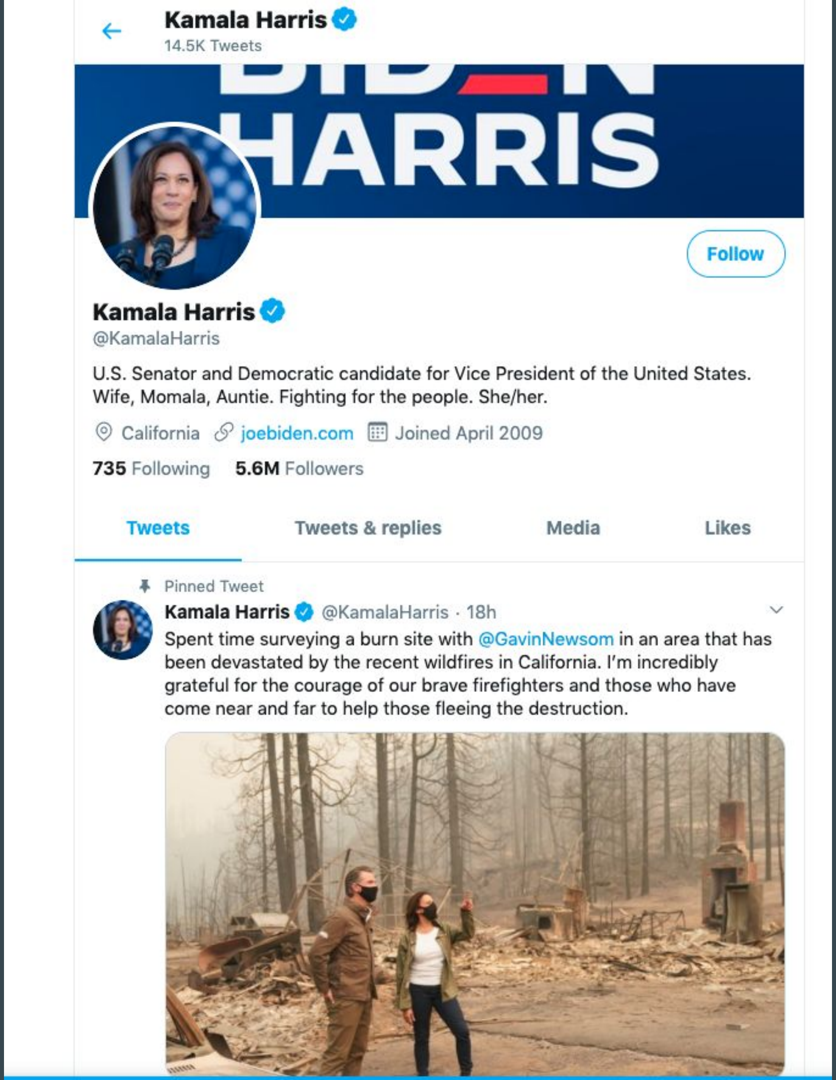
Language
en

Keyword

Date created
2020

Type of Work

Social media on the web



Kamala Harris
14.5K Tweets

Kamala Harris
@KamalaHarris

U.S. Senator and Democratic candidate for Vice President of the United States. Wife, Momala, Auntie. Fighting for the people. She/her.

California joebiden.com Joined April 2009

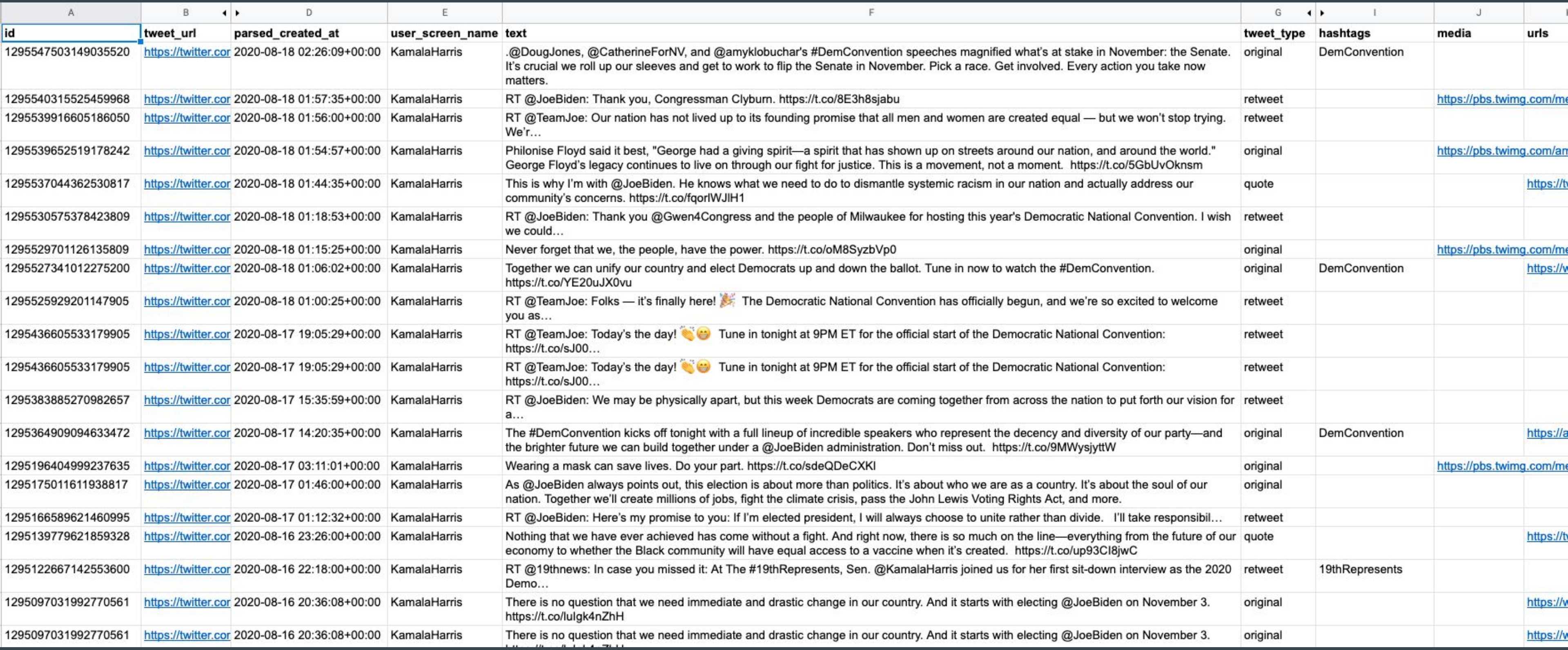
735 Following 5.6M Followers

Pinned Tweet

Kamala Harris @KamalaHarris · 18h

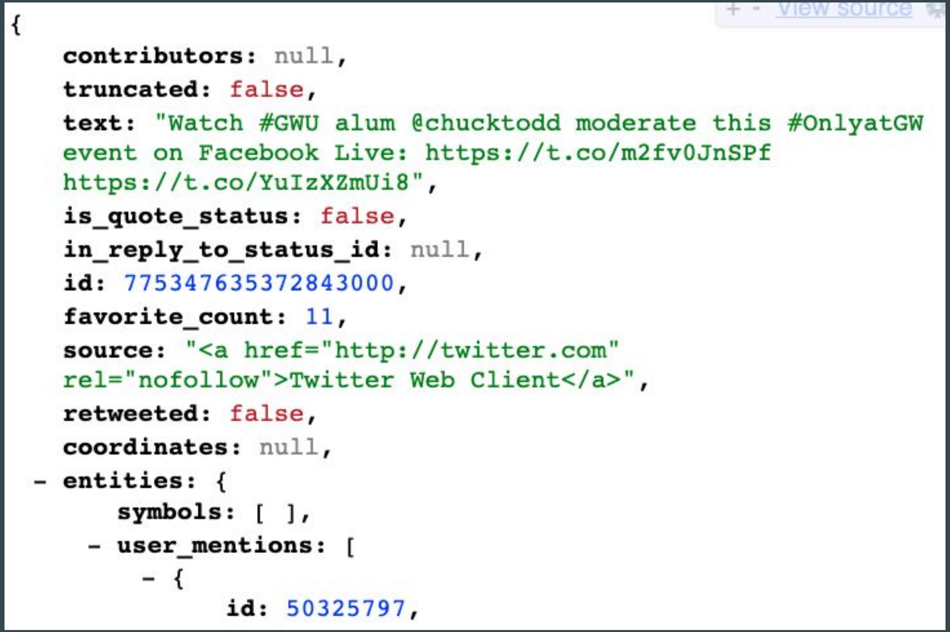
Spent time surveying a burn site with @GavinNewsom in an area that has been devastated by the recent wildfires in California. I'm incredibly grateful for the courage of our brave firefighters and those who have come near and far to help those fleeing the destruction.

Social media as data



A	B	D	E	F	G	I	J	
id	tweet_url	parsed_created_at	user_screen_name	text	tweet_type	hashtags	media	urls
1295547503149035520	https://twitter.cor	2020-08-18 02:26:09+00:00	KamalaHarris	.@DougJones, @CatherineForNV, and @amyklobuchar's #DemConvention speeches magnified what's at stake in November: the Senate. It's crucial we roll up our sleeves and get to work to flip the Senate in November. Pick a race. Get involved. Every action you take now matters.	original	DemConvention		
1295540315525459968	https://twitter.cor	2020-08-18 01:57:35+00:00	KamalaHarris	RT @JoeBiden: Thank you, Congressman Clyburn. https://t.co/8E3h8sjabu	retweet		https://pbs.twimg.com/me	
1295539916605186050	https://twitter.cor	2020-08-18 01:56:00+00:00	KamalaHarris	RT @TeamJoe: Our nation has not lived up to its founding promise that all men and women are created equal — but we won't stop trying. We'r…	retweet			
1295539652519178242	https://twitter.cor	2020-08-18 01:54:57+00:00	KamalaHarris	Philonise Floyd said it best, "George had a giving spirit—a spirit that has shown up on streets around our nation, and around the world." George Floyd's legacy continues to live on through our fight for justice. This is a movement, not a moment. https://t.co/5GbUvOknsm	original		https://pbs.twimg.com/am	
1295537044362530817	https://twitter.cor	2020-08-18 01:44:35+00:00	KamalaHarris	This is why I'm with @JoeBiden. He knows what we need to do to dismantle systemic racism in our nation and actually address our community's concerns. https://t.co/fqorlWJlH1	quote			https://t
1295530575378423809	https://twitter.cor	2020-08-18 01:18:53+00:00	KamalaHarris	RT @JoeBiden: Thank you @Gwen4Congress and the people of Milwaukee for hosting this year's Democratic National Convention. I wish we could…	retweet			
1295529701126135809	https://twitter.cor	2020-08-18 01:15:25+00:00	KamalaHarris	Never forget that we, the people, have the power. https://t.co/oM8SyzbVp0	original		https://pbs.twimg.com/me	
1295527341012275200	https://twitter.cor	2020-08-18 01:06:02+00:00	KamalaHarris	Together we can unify our country and elect Democrats up and down the ballot. Tune in now to watch the #DemConvention. https://t.co/YE20uJX0vu	original	DemConvention		https://w
1295525929201147905	https://twitter.cor	2020-08-18 01:00:25+00:00	KamalaHarris	RT @TeamJoe: Folks — it's finally here! 🎉 The Democratic National Convention has officially begun, and we're so excited to welcome you as…	retweet			
1295436605533179905	https://twitter.cor	2020-08-17 19:05:29+00:00	KamalaHarris	RT @TeamJoe: Today's the day! 👏😁 Tune in tonight at 9PM ET for the official start of the Democratic National Convention: https://t.co/sJ00…	retweet			
1295436605533179905	https://twitter.cor	2020-08-17 19:05:29+00:00	KamalaHarris	RT @TeamJoe: Today's the day! 👏😁 Tune in tonight at 9PM ET for the official start of the Democratic National Convention: https://t.co/sJ00…	retweet			
1295383885270982657	https://twitter.cor	2020-08-17 15:35:59+00:00	KamalaHarris	RT @JoeBiden: We may be physically apart, but this week Democrats are coming together from across the nation to put forth our vision for a…	retweet			
1295364909094633472	https://twitter.cor	2020-08-17 14:20:35+00:00	KamalaHarris	The #DemConvention kicks off tonight with a full lineup of incredible speakers who represent the decency and diversity of our party—and the brighter future we can build together under a @JoeBiden administration. Don't miss out. https://t.co/9MWysjyttW	original	DemConvention		https://a
1295196404999237635	https://twitter.cor	2020-08-17 03:11:01+00:00	KamalaHarris	Wearing a mask can save lives. Do your part. https://t.co/sdeQDeCXKl	original		https://pbs.twimg.com/me	
1295175011611938817	https://twitter.cor	2020-08-17 01:46:00+00:00	KamalaHarris	As @JoeBiden always points out, this election is about more than politics. It's about who we are as a country. It's about the soul of our nation. Together we'll create millions of jobs, fight the climate crisis, pass the John Lewis Voting Rights Act, and more.	original			
1295166589621460995	https://twitter.cor	2020-08-17 01:12:32+00:00	KamalaHarris	RT @JoeBiden: Here's my promise to you: If I'm elected president, I will always choose to unite rather than divide. I'll take responsibil…	retweet			
1295139779621859328	https://twitter.cor	2020-08-16 23:26:00+00:00	KamalaHarris	Nothing that we have ever achieved has come without a fight. And right now, there is so much on the line—everything from the future of our economy to whether the Black community will have equal access to a vaccine when it's created. https://t.co/up93CI8jwC	quote			https://t
1295122667142553600	https://twitter.cor	2020-08-16 22:18:00+00:00	KamalaHarris	RT @19thnews: In case you missed it: At The #19thRepresents, Sen. @KamalaHarris joined us for her first sit-down interview as the 2020 Demo…	retweet	19thRepresents		
1295097031992770561	https://twitter.cor	2020-08-16 20:36:08+00:00	KamalaHarris	There is no question that we need immediate and drastic change in our country. And it starts with electing @JoeBiden on November 3. https://t.co/lulgk4nZhH	original			https://w
1295097031992770561	https://twitter.cor	2020-08-16 20:36:08+00:00	KamalaHarris	There is no question that we need immediate and drastic change in our country. And it starts with electing @JoeBiden on November 3.	original			https://w

Tweets are data, too



```
{
    contributors: null,
    truncated: false,
    text: "Watch #GWU alum @chucktodd moderate this #OnlyatGW
    event on Facebook Live: https://t.co/m2fv0JnSPf
    https://t.co/YuIzXZmUi8",
    is_quote_status: false,
    in_reply_to_status_id: null,
    id: 775347635372843000,
    favorite_count: 11,
    source: "<a href="http://twitter.com"
    rel="nofollow">Twitter Web Client</a>",
    retweeted: false,
    coordinates: null,
  - entities: {
        symbols: [ ],
      - user_mentions: [
          - {
                id: 50325797,
```

Twitter's guide to the structure of a tweet

JSON: JavaScript Object Notation

- `{ key: value, key: value… }`
- keys are strings
- values may be:
    - string: in quotes: `"GW"`
    - number
    - boolean - `true` or `false`
    - another JSON object
    - array (denoted by square brackets `[ ]`) of JSON objects
    - `null`

JSON example

```
{
   "text": "Yesterday, #GWU students, faculty,
staff...https://t.co/8Tz29odc11",
   "favorite_count": 56,
   "truncated": false,
   "entities": {
      "user_mentions": [],
      "hashtags": {
         "indices": [11, 15],
         "text": "GWU"
     }
   }
}
```

Social Media APIs

What's an API?

“Application Programming Interface”

Allows you to request or send data to another service on the web, using HTTP.

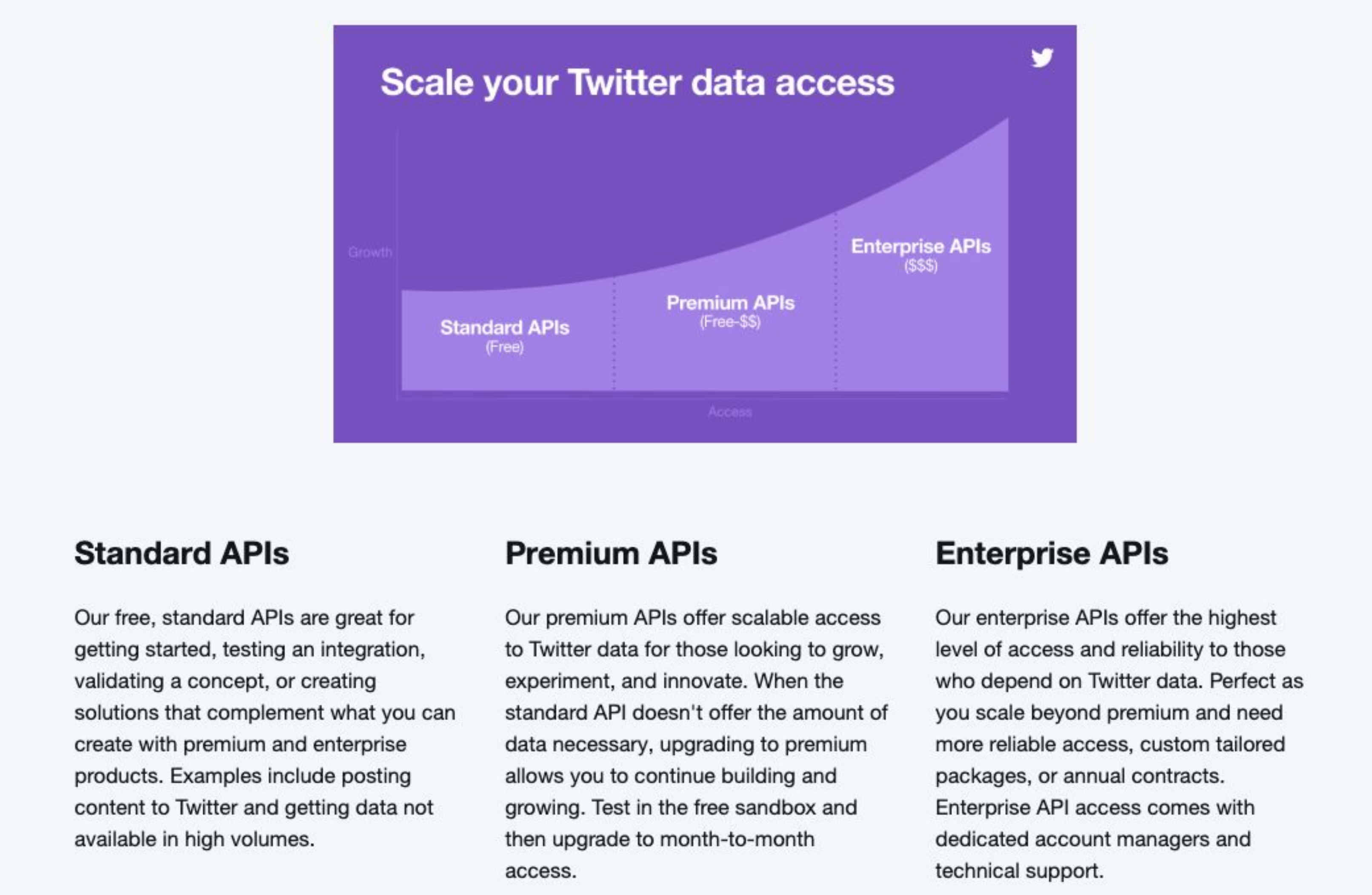
Request: http://an.api.com/query?term=pizza

Response: structured data (XML, JSON)

Why use an API for working with social media?

- Data you can't get by "scraping" the web.
- Data is in a structured format, easier for analyzing.

The Twitter API



Standard APIs

Our free, standard APIs are great for getting started, testing an integration, validating a concept, or creating solutions that complement what you can create with premium and enterprise products. Examples include posting content to Twitter and getting data not available in high volumes.

Premium APIs

Our premium APIs offer scalable access to Twitter data for those looking to grow, experiment, and innovate. When the standard API doesn't offer the amount of data necessary, upgrading to premium allows you to continue building and growing. Test in the free sandbox and then upgrade to month-to-month access.

Enterprise APIs

Our enterprise APIs offer the highest level of access and reliability to those who depend on Twitter data. Perfect as you scale beyond premium and need more reliable access, custom tailored packages, or annual contracts. Enterprise API access comes with dedicated account managers and technical support.

Understanding the Twitter API

- There are many Twitter APIs, only some are free.
- Their restrictions and affordances shape what you can collect.
- Understanding the APIs allows you to best choose which research questions can be addressed.

Most useful API methods for collecting tweets

- User timeline: GET statuses/user_timeline
  - Up to the most recent 3,200 tweets
- Search: GET search/tweets
  - Sampling of tweets from last 7 days.
  - Query by keyword, phrases, hashtags, author, date, more.
  - Not the same as search via twitter.com
- Filter stream: POST statuses/filter
  - Filter by keyword, user, or location

User timeline: GET statuses/user_timeline

- Gets most recent tweets posted by a user.
- Limited to last 3,200 tweets.
- Returns 200 at a time, so must page.
- Rate limit: 900 tweets per 15 minutes
- `https://api.twitter.com/1.1/statuses/user_timeline.json?screen_name=gelmanlibrary&max_id=829886156334571 5`

Search: GET search/tweets

- Search recent tweets.
  - Sampling of tweets from last 7 days.
  - Query by keyword, phrases, hashtags, author, date, more.
- Returns up to 100 at a time, so must page.
- Not the same as search on Twitter website.
- Rate limit: 180 tweets per 15 minutes
- `https://api.twitter.com/1.1/search/tweets.json?q=%23onlyatgw`

Filter Stream: POST statuses/filter

- Real-time filtering of all public tweets.
  - Filter by keyword, user, or location.
- Continue to receive additional tweets over a single call to API. (No paging.)
- Limits:
  - When high volume, will not receive all tweets.
  - One stream at a time per set of credentials.
- `https://stream.twitter.com/1.1/statuses/filter.json?track=gwu`

More Twitter API methods

Get a specific tweet: GET users/lookup

Get user info: GET users/lookup

Get trends near a location: GET trends/place

More: developer.twitter.com/en/docs

Acquiring Twitter datasets

Options for acquiring a Twitter dataset

- Collect a new dataset.
- Use an existing dataset.
- Purchase data or access to a platform.

Collecting new Twitter data

Collecting a new dataset - using coding

Command line:

- Twarc: github.com/docnow/twarc
- Twurl: github.com/twitter/twurl

Python libraries

- twarc github.com/DocNow/twarc
- tweepy: www.tweepy.org

R package: `rtweet` - github.com/ropensci/rtweet

Collecting a new dataset - no coding required

- Social Feed Manager: go.gwu.edu/sfmgw
- TAGS (Twitter Archiving Google Sheet): tags.hawksey.info

Social Feed Manager software

- Open source software by GW Libraries.
- User interface for collecting, managing, and exporting social media data from APIs.
- Collect from Twitter, Tumblr, Flickr, Sina Weibo.
- Libraries run this for their community as a service.

More: go.gwu.edu/sfm

Hands-on: Social Feed Manager

Steps we'll perform:

1. Sign up
2. Request credentials (API keys)
3. Create a collection
4. Perform a harvest
5. Export data

Go to: gwsfm-sandbox.wrlc.org

Exporting datasets

- Formats: Excel, CSV, JSON
- Limit by date ranges
- Splits into separate files

Using existing Twitter data

Datasets from other researchers

- Twitter's terms generally do not allow datasets of full JSON data to be shared.
- OK to share: Text file of tweet identifiers

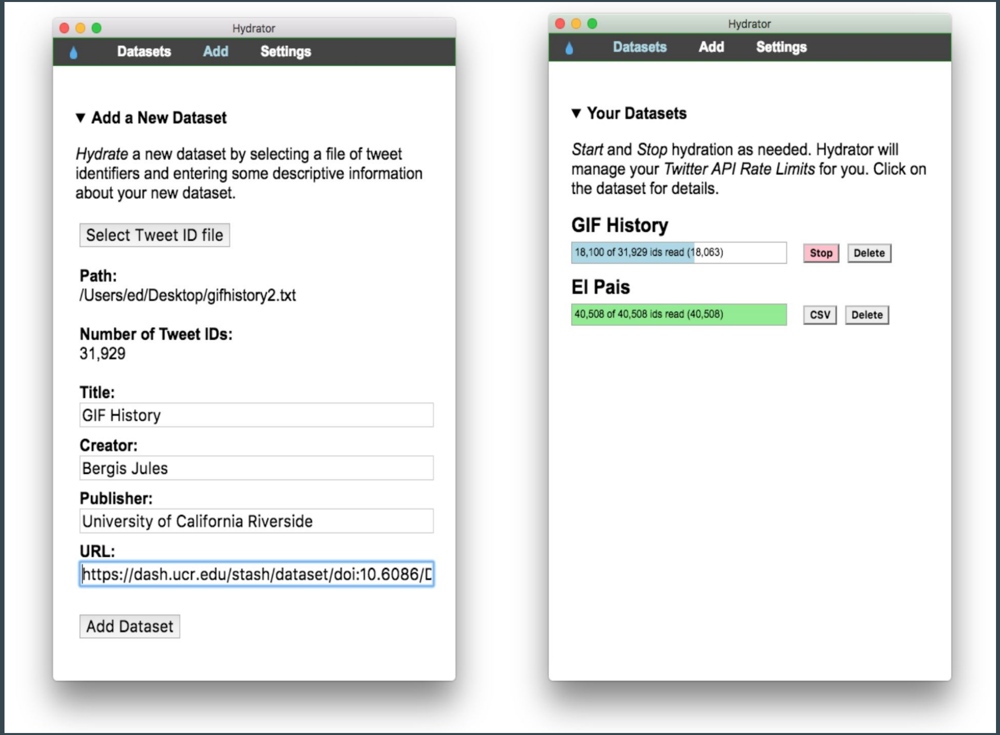
```
id_str: "775347040196894720"
```

- Use Twitter API to request tweets by identifier and get back the full tweet.
- Won't include deleted/protected tweets.

Working with tweet identifiers

Hydrator desktop app

https://github.com/DocNow/hydrator



▼ Add a New Dataset

Hydrate a new dataset by selecting a file of tweet identifiers and entering some descriptive information about your new dataset.

Path:
/Users/ed/Desktop/gifhistory2.txt

Number of Tweet IDs:
31,929

Title:
GIF History

Creator:
Bergis Jules

Publisher:
University of California Riverside

URL:
https://dash.ucr.edu/stash/dataset/doi:10.6086/D

▼ Your Datasets

Start and *Stop* hydration as needed. Hydrator will manage your *Twitter API Rate Limits* for you. Click on the dataset for details.

GIF History

18,100 of 31,929 ids read (18,063)

El Pais

40,508 of 40,508 ids read (40,508)

Using an existing dataset

- DocNow Catalog: www.docnow.io/catalog/
- Data repositories such as Dataverse
- TweetSets: tweetsets.library.gwu.edu
    - Datasets collected by GW Libraries.
    - Full tweets available as JSON or CSV
    - Only for GW users, for academic purposes only.

Datasets collected by GW Libraries

- Coronavirus
- 2020 U.S. Presidential election
- 2018 U.S. midterm election
- 2016 U.S. election (280 million tweets)
- Congress (all senators and representatives)
- Federal govt (3000 U.S. government accounts)
- News outlets (4500 media organization accounts)
- Hurricanes
- Climate change

More ...

TweetSets

Steps we'll demo:

1. Select a source dataset.
2. Filter the source dataset.
3. Create a new dataset.
4. Generate and download dataset derivatives.

tweetsets.library.gwu.edu

Purchasing data or access to a platform

Options

- Subscribe to an analytics platform such as CrimsonHexagon. Note limitations on data export.
- Subscribe to Twitter Premium or Enterprise APIs.
- Purchase historical batch data from Twitter.
- Subscribe to historical search API access from Twitter.
- Free: Twitter academic research product track

Can I get Tweets from the past without cost?

- If GW collected it already: yes (TweetSets or SFM)
- If someone else collected it:
  - Need to hydrate tweet IDs, won’t be all tweets.
- Using Twitter collections in SFM:
  - User timeline: up to ~3,200 tweets per account
  - Search: ~7 days
  - Filter: No.
- Via the Academic Product Track (faculty and grad students can apply). Requires using command-line tools.

FAQ: Are tweets geotagged?

- Geotagging is opt-in. Only ~2% geotagged.

  Lat, long or place name (e.g., DC or Middle Earth)

- Search API: Limit to a specified distance from a point.
- Filter Stream: Limit to a bounding box.

More: gwu-libraries.github.io/sfm-ui/posts/2017-04-12-geographic-collecting

Exploring and analyzing Twitter data

Before analysis

Clean and validate your data.

- Are the terms you queried used for other meanings and events?
- Are the accounts valid?
- Are there gaps in the data?

Working with datasets

- Jupyter notebooks for Python and pandas analysis: bit.ly/2uhN252 also see here
- R
- jq command-line tool

“Recipes for Twitter data” bit.ly/2t9cStF

- Excel or Google Sheets

Other social media platforms

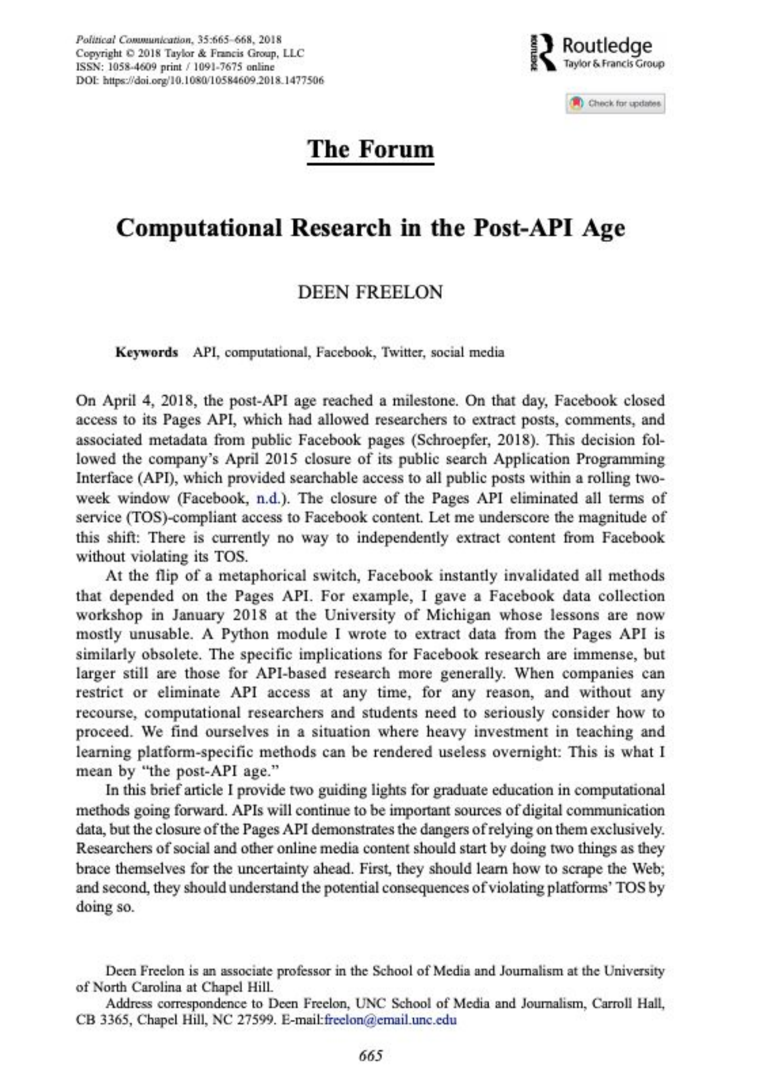
Services for datasets

CrowdTangle: must apply for access for research purposes. Faculty and PhD students only

- Facebook: 6M+ Facebook pages, groups, and verified profiles. This includes all public Facebook pages and groups with more than 100K likes (automated via API), all US-based public groups with 2k+ members, and all verified profiles.
- Instagram: 2M+ public Instagram accounts. This includes all public Instagram accounts with more than 75K followers, as well as all verified accounts.
- Reddit: ~20K+ of the most active sub-reddits. Built and maintained in

Access for other platforms?

- Facebook: Graph API provides some data. Try Facepager application.
- Instagram: no API available anymore. Web scrape with command-line tools Instagram Scraper or Instaloader.
- YouTube: Data API available for metadata and comments. Can incur costs. Consider youtube-dl.org for downloading videos.
- Reddit: API provides access to data via code.



Political Communication, 35:665–668, 2018
Copyright © 2018 Taylor & Francis Group, LLC
ISSN: 1058-4609 print / 1091-7675 online
DOI: https://doi.org/10.1080/10584609.2018.1477506

The Forum

Computational Research in the Post-API Age

DEEN FREELON

Keywords API, computational, Facebook, Twitter, social media

On April 4, 2018, the post-API age reached a milestone. On that day, Facebook closed access to its Pages API, which had allowed researchers to extract posts, comments, and associated metadata from public Facebook pages (Schroepfer, 2018). This decision followed the company's April 2015 closure of its public search Application Programming Interface (API), which provided searchable access to all public posts within a rolling two-week window (Facebook, n.d.). The closure of the Pages API eliminated all terms of service (TOS)-compliant access to Facebook content. Let me underscore the magnitude of this shift: There is currently no way to independently extract content from Facebook without violating its TOS.

At the flip of a metaphorical switch, Facebook instantly invalidated all methods that depended on the Pages API. For example, I gave a Facebook data collection workshop in January 2018 at the University of Michigan whose lessons are now mostly unusable. A Python module I wrote to extract data from the Pages API is similarly obsolete. The specific implications for Facebook research are immense, but larger still are those for API-based research more generally. When companies can restrict or eliminate API access at any time, for any reason, and without any recourse, computational researchers and students need to seriously consider how to proceed. We find ourselves in a situation where heavy investment in teaching and learning platform-specific methods can be rendered useless overnight: This is what I mean by "the post-API age."

In this brief article I provide two guiding lights for graduate education in computational methods going forward. APIs will continue to be important sources of digital communication data, but the closure of the Pages API demonstrates the dangers of relying on them exclusively. Researchers of social and other online media content should start by doing two things as they brace themselves for the uncertainty ahead. First, they should learn how to scrape the Web; and second, they should understand the potential consequences of violating platforms' TOS by doing so.

Deen Freelon is an associate professor in the School of Media and Journalism at the University of North Carolina at Chapel Hill.

Address correspondence to Deen Freelon, UNC School of Media and Journalism, Carroll Hall, CB 3365, Chapel Hill, NC 27599. E-mail:freelon@email.unc.edu

What do you do when there's no API available?

Web scraping and capture tools are an alternative approach.

Deen Freelon (2018) [Computational Research in the Post-API Age](), *Political Communication*, 35:4, 665-668. [Also available as preprint.]()

Conifer (formerly Webrecorder)

conifer.rhizome.org

- “Record” your web browsing and capture sites as viewed by a human.
- Provides a complementary view to API data.
- Sign in for 5GB account. Can make collections public or export them.

Ethical considerations

Social media data comes from people

- Consider impact of your work on the creator of the social media.
- Do not have creator's permission for research.
- Impact on creator is balanced against public good of your research.
- Requires judgment call.

More: go.gwu.edu/sfmethics

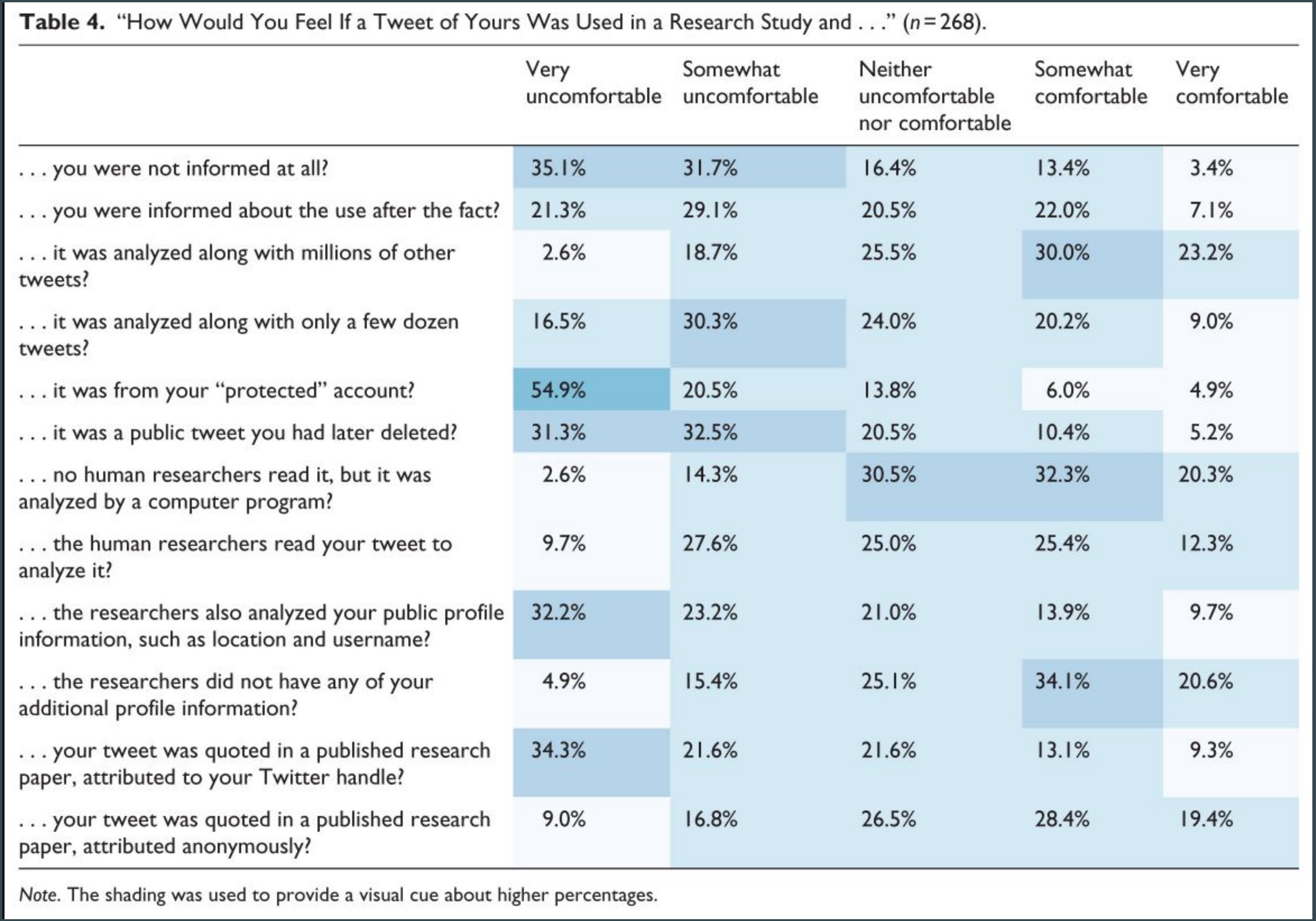


Table 4. "How Would You Feel If a Tweet of Yours Was Used in a Research Study and . . ." ($n = 268$).

	Very uncomfortable	Somewhat uncomfortable	Neither uncomfortable nor comfortable	Somewhat comfortable	Very comfortable
. . . you were not informed at all?	35.1%	31.7%	16.4%	13.4%	3.4%
. . . you were informed about the use after the fact?	21.3%	29.1%	20.5%	22.0%	7.1%
. . . it was analyzed along with millions of other tweets?	2.6%	18.7%	25.5%	30.0%	23.2%
. . . it was analyzed along with only a few dozen tweets?	16.5%	30.3%	24.0%	20.2%	9.0%
. . . it was from your "protected" account?	54.9%	20.5%	13.8%	6.0%	4.9%
. . . it was a public tweet you had later deleted?	31.3%	32.5%	20.5%	10.4%	5.2%
. . . no human researchers read it, but it was analyzed by a computer program?	2.6%	14.3%	30.5%	32.3%	20.3%
. . . the human researchers read your tweet to analyze it?	9.7%	27.6%	25.0%	25.4%	12.3%
. . . the researchers also analyzed your public profile information, such as location and username?	32.2%	23.2%	21.0%	13.9%	9.7%
. . . the researchers did not have any of your additional profile information?	4.9%	15.4%	25.1%	34.1%	20.6%
. . . your tweet was quoted in a published research paper, attributed to your Twitter handle?	34.3%	21.6%	21.6%	13.1%	9.3%
. . . your tweet was quoted in a published research paper, attributed anonymously?	9.0%	16.8%	26.5%	28.4%	19.4%

Note. The shading was used to provide a visual cue about higher percentages.

"Participant Perceptions of Twitter Research Ethics." Casey Fiesler, Nicholas Proferes, *Social Media + Society*. First published March 10, 2018.

Data collecting

Be thoughtful collecting social media of:

- Vulnerable individuals (e.g., minors, social activists)
- Sensitive or harmful topics (e.g., questionable behavior, mental illness)
- Geography-based collecting

Data analysis

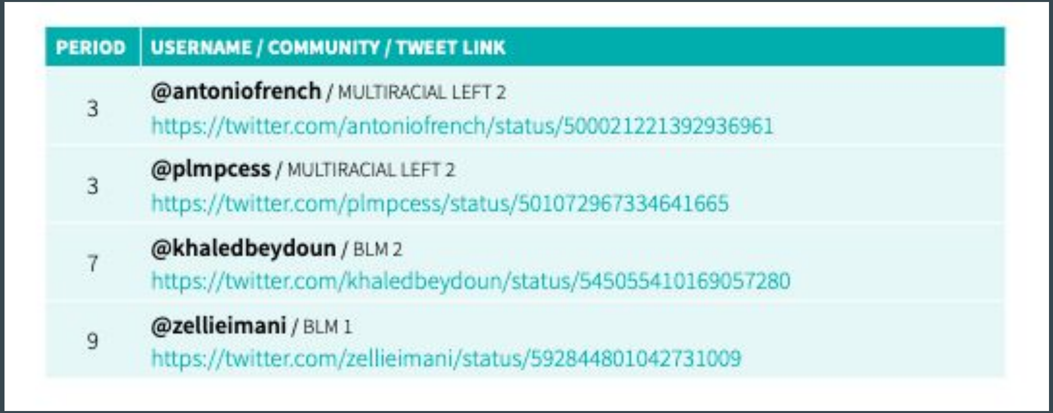
- **Inferring** individual characteristics:
  - Health (including pregnancy)
  - Negative financial status or condition
  - Political affiliation or beliefs
  - Racial or ethnic origin
  - Religious or philosophical affiliation or beliefs
  - Sex life or sexual orientation
  - Trade union membership
  - Alleged or actual commission of a crime
- Off-Twitter matching
- Surveillance
- Facial recognition

Twitter's restricted use cases

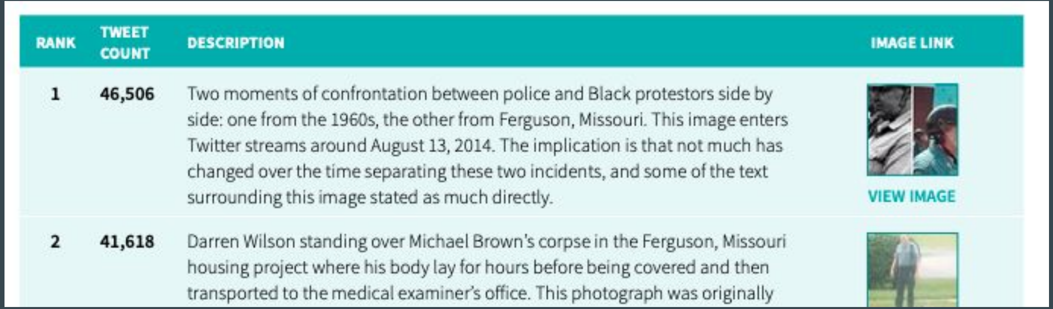
Publishing

- When possible, get permission from creator for quotes.
- Do not rely on anonymizing posts.
- Link to a specific tweet rather than republish.
- Include your ethical decision-making in your paper.

BEYOND THE
HASHTAGS
#Ferguson, #Blacklivesmatter, and the online struggle for offline justice
DEEN FREELON
CHARLTON D. MCILWAIN
MEREDITH D. CLARK

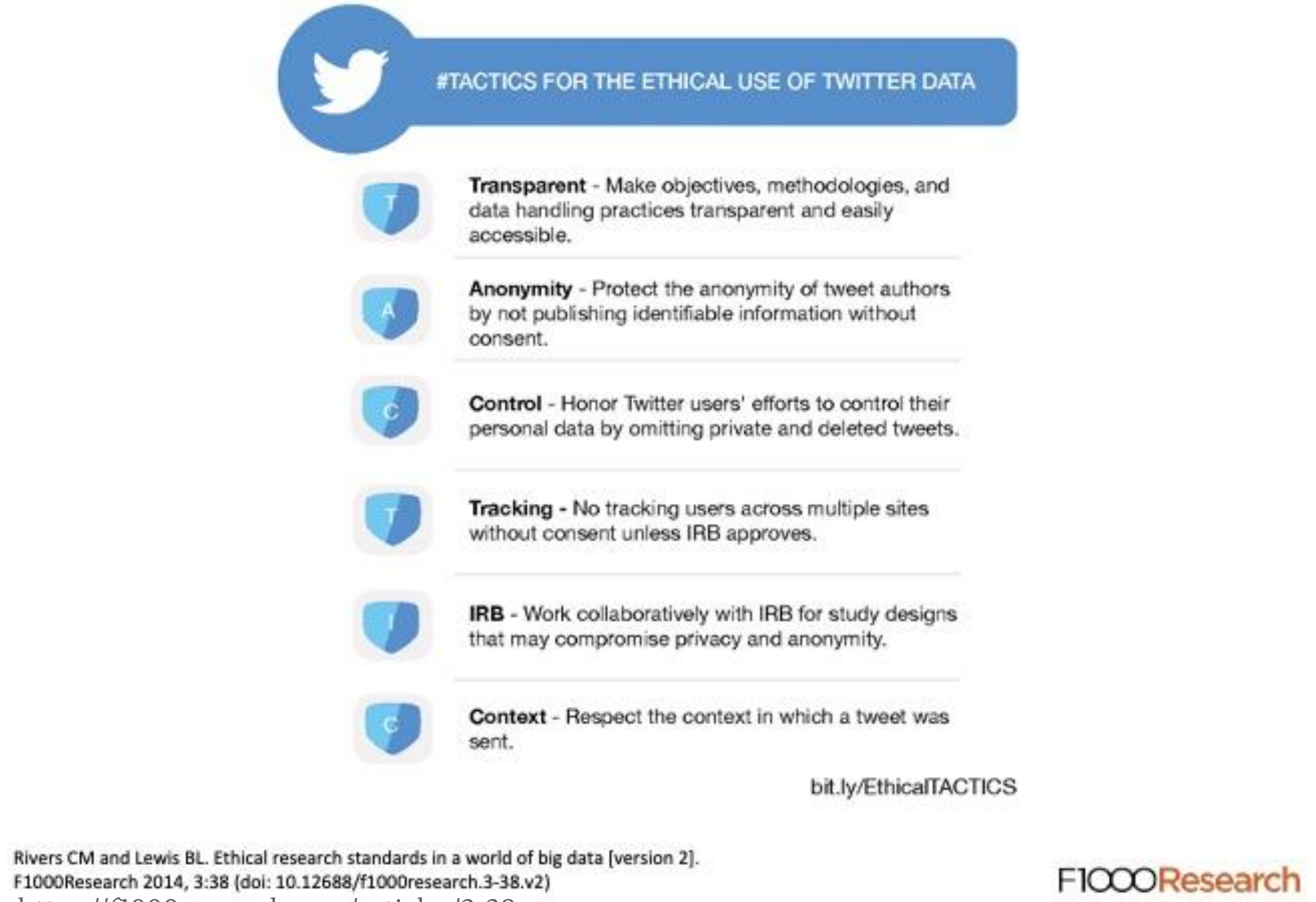


PERIOD	USERNAME / COMMUNITY / TWEET LINK
3	**@antoniofrench** / MULTIRACIAL LEFT 2 https://twitter.com/antoniofrench/status/500021221392936961
3	**@plmpcess** / MULTIRACIAL LEFT 2 https://twitter.com/plmpcess/status/501072967334641665
7	**@khaledbeydoun** / BLM 2 https://twitter.com/khaledbeydoun/status/545055410169057280
9	**@zellieimani** / BLM 1 https://twitter.com/zellieimani/status/592844801042731009



RANK	TWEET COUNT	DESCRIPTION	IMAGE LINK
1	46,506	Two moments of confrontation between police and Black protestors side by side: one from the 1960s, the other from Ferguson, Missouri. This image enters Twitter streams around August 13, 2014. The implication is that not much has changed over the time separating these two incidents, and some of the text surrounding this image stated as much directly.	VIEW IMAGE
2	41,618	Darren Wilson standing over Michael Brown's corpse in the Ferguson, Missouri housing project where his body lay for hours before being covered and then transported to the medical examiner's office. This photograph was originally	

Freelon, Deen and McIlwain, Charlton D. and Clark, Meredith, Beyond the Hashtags: #Ferguson, #Blacklivesmatter, and the Online Struggle for Offline Justice (February 29, 2016). Center for Media & Social Impact, American University, Forthcoming, Available at SSRN: https://ssrn.com/abstract=2747066 or http://dx.doi.org/10.2139/ssrn.2747066



#TACTICS FOR THE ETHICAL USE OF TWITTER DATA

Transparent - Make objectives, methodologies, and data handling practices transparent and easily accessible.

Anonymity - Protect the anonymity of tweet authors by not publishing identifiable information without consent.

Control - Honor Twitter users' efforts to control their personal data by omitting private and deleted tweets.

Tracking - No tracking users across multiple sites without consent unless IRB approves.

IRB - Work collaboratively with IRB for study designs that may compromise privacy and anonymity.

Context - Respect the context in which a tweet was sent.

bit.ly/EthicalTACTICS

Rivers CM and Lewis BL. Ethical research standards in a world of big data [version 2]. F1000Research 2014, 3:38 (doi: 10.12688/f1000research.3-38.v2)

F1000Research

https://f1000research.com/articles/3-38

Data sharing

- Get familiar with platform terms of use.
  - Don’t republish full datasets
  - Share in accordance with terms (e.g., tweet ids only)
  - Consider copyright
- Sharing summary statistics is usually OK.

Questions?

Make a consultation appointment:
calendly.com/social-media-consulting-gw

Social Feed Manager team:
sfm@gwu.edu

Laura Wrubel	Dan Kerchner
lwrubel@gwu.edu	kerchner@gwu.edu