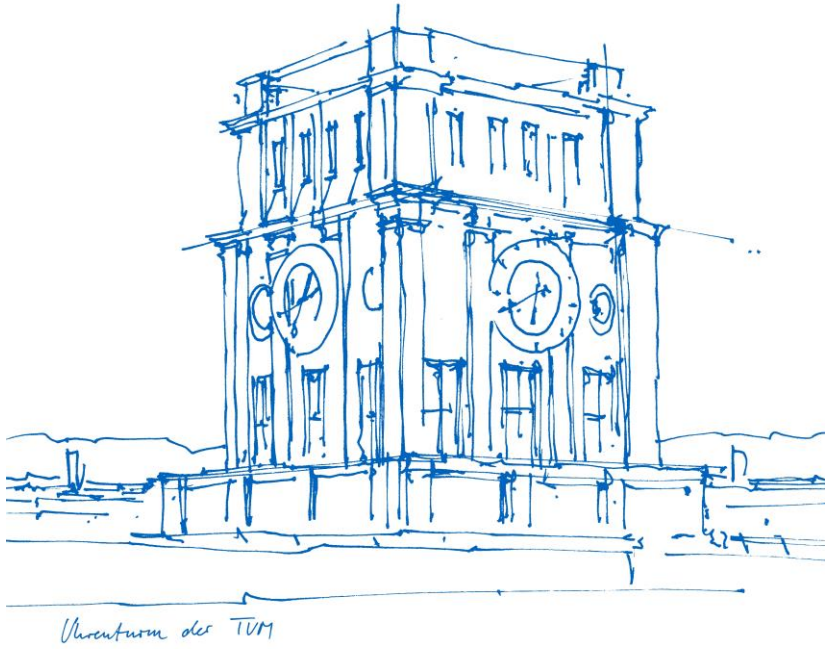
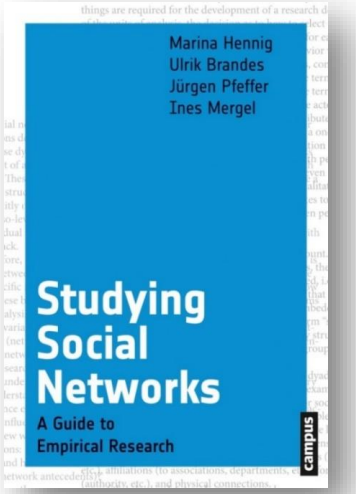
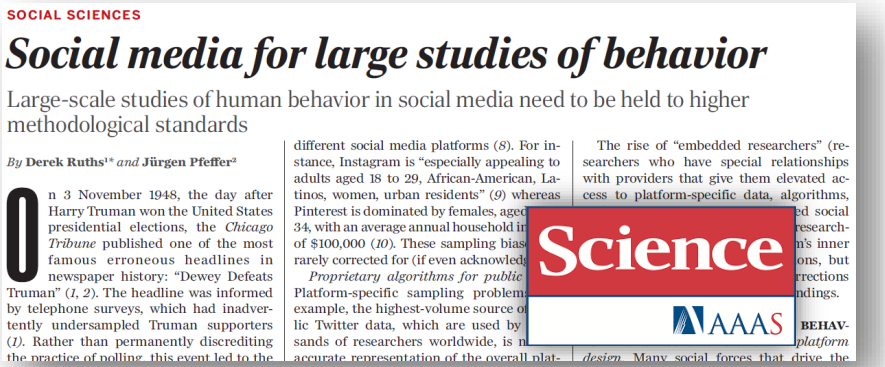
It’s a new day…
Life after the Twitter API

Prof. Dr. Jürgen Pfeffer
Computational Social Science
Technical University of Munich
School of Social Sciences & Technology
@JurgenPfeffer






SOCIAL SCIENCES

Social media for large studies of behavior

Large-scale studies of human behavior in social media need to be held to higher methodological standards

By **Derek Ruths**[1]* *and* **Jürgen Pfeffer**[2]

On 3 November 1948, the day after Harry Truman won the United States presidential elections, the *Chicago Tribune* published one of the most famous erroneous headlines in newspaper history: "Dewey Defeats Truman" (*1*, *2*). The headline was informed by telephone surveys, which had inadvertently undersampled Truman supporters (*1*). Rather than permanently discrediting the practice of polling, this event led to the

different social media platforms (*8*). For instance, Instagram is "especially appealing to adults aged 18 to 29, African-American, Latinos, women, urban residents" (*9*) whereas Pinterest is dominated by females, aged ed social 34, with an average annual household in research- of $100,000 (*10*). These sampling bias n's inner rarely corrected for (if even acknowled ons, but

Proprietary algorithms for public rrections Platform-specific sampling problem ndings. example, the highest-volume source o lic Twitter data, which are used by BEHAV- sands of researchers worldwide, is n *platform* accurate representation of the overall plat- *design*. Many social forces that drive the

Science

AAAS

Vita

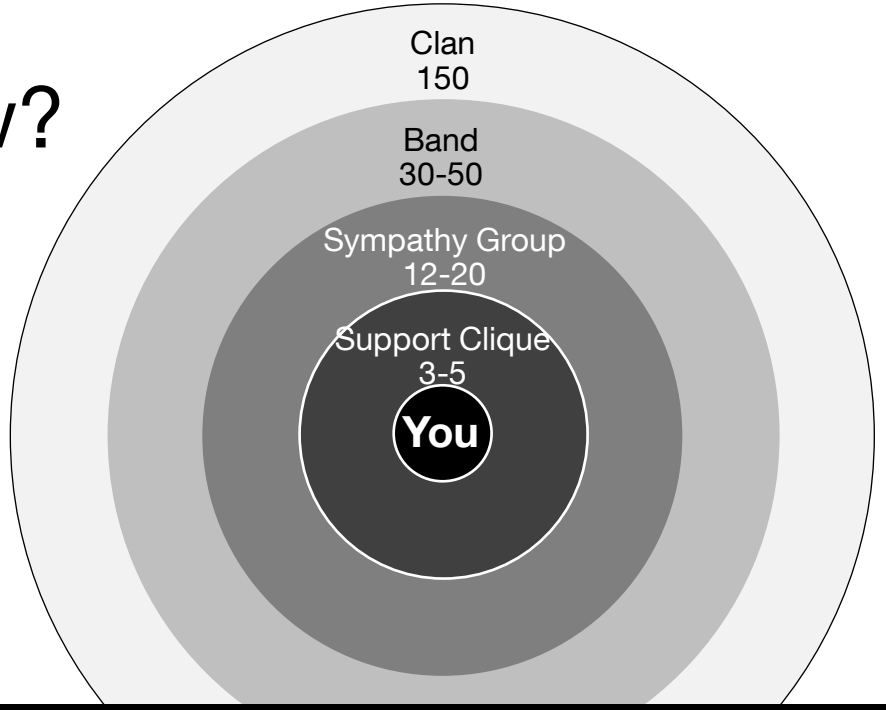
- BA Computer Science, Ph.D. Business Informatics, Vienna University of Technology
- 10 years consulting and non-university research
- 2012-2015: Assistant Professor @ Carnegie Mellon University in Pittsburgh
- 2016- : Professor of Computational Social Science @ TU Munich

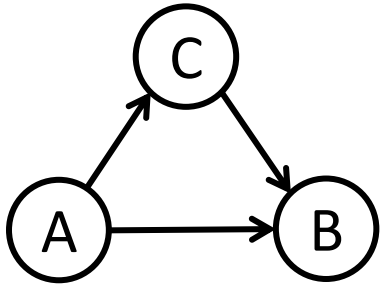
Research focus

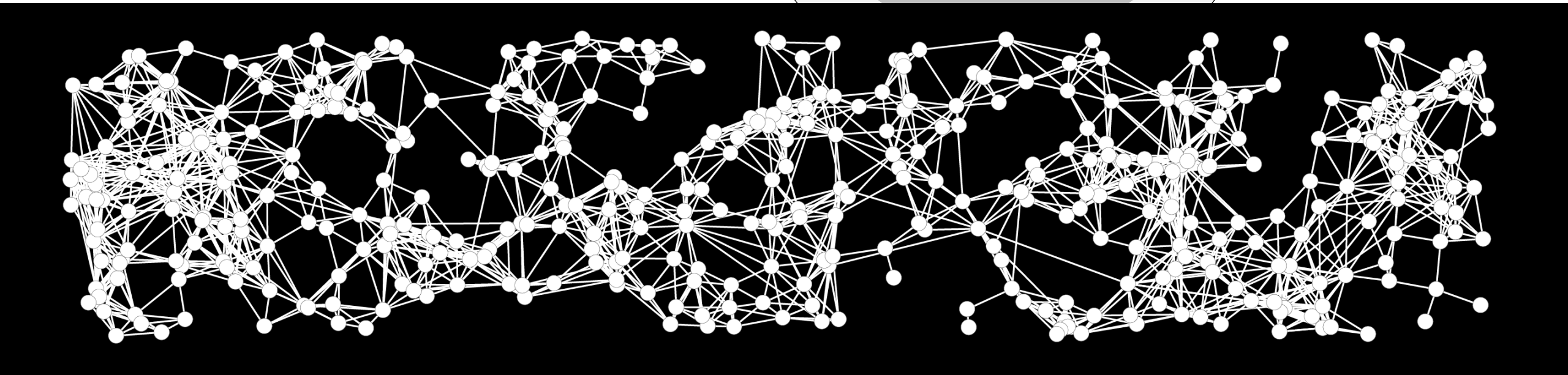
- Computational analysis of organizations and societies
- Special emphasis on large-scale systems, e.g., social media
- Methodological and algorithmic challenges
- Network analysis methods
- Information Visualization

What do I really want to know?

I want to understand the structure and dynamics of the social world!







Motivation - Computational Social Science

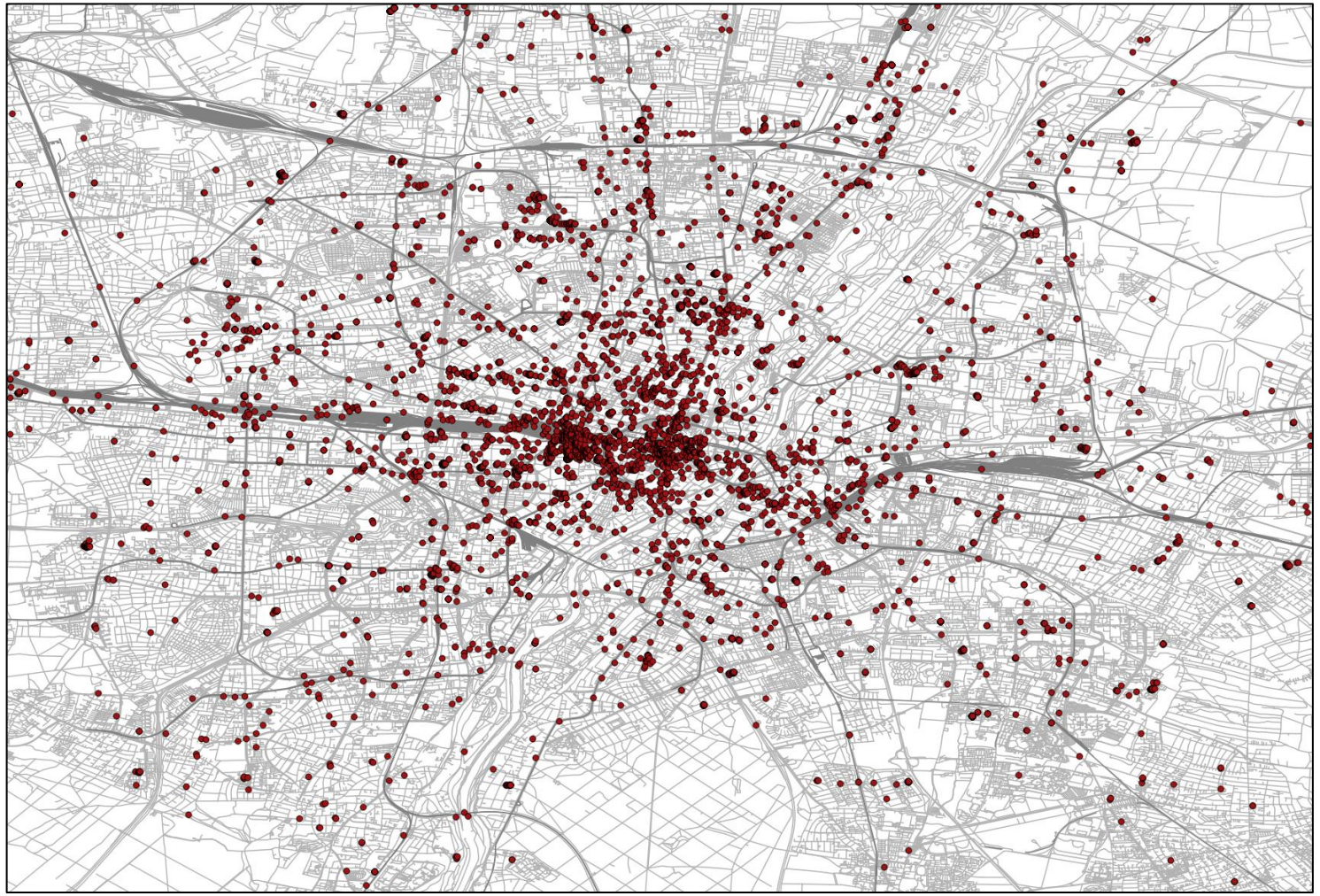
"…access to terabytes of **data describing minute-by-minute interactions** and locations of entire populations of individuals… [will] offer qualitatively new perspectives on **collective human behavior**."

Lazer, D. et al. (2009). Computational social science. Science, 323, 721-723.

"… refers to the academic sub-disciplines concerned with **computational approaches** to the social sciences. This means that computers are used to model, simulate, and analyze **social phenomena**."

Wikipedia, 3/23/2022

Data Hopes and Dreams



Research Questions?

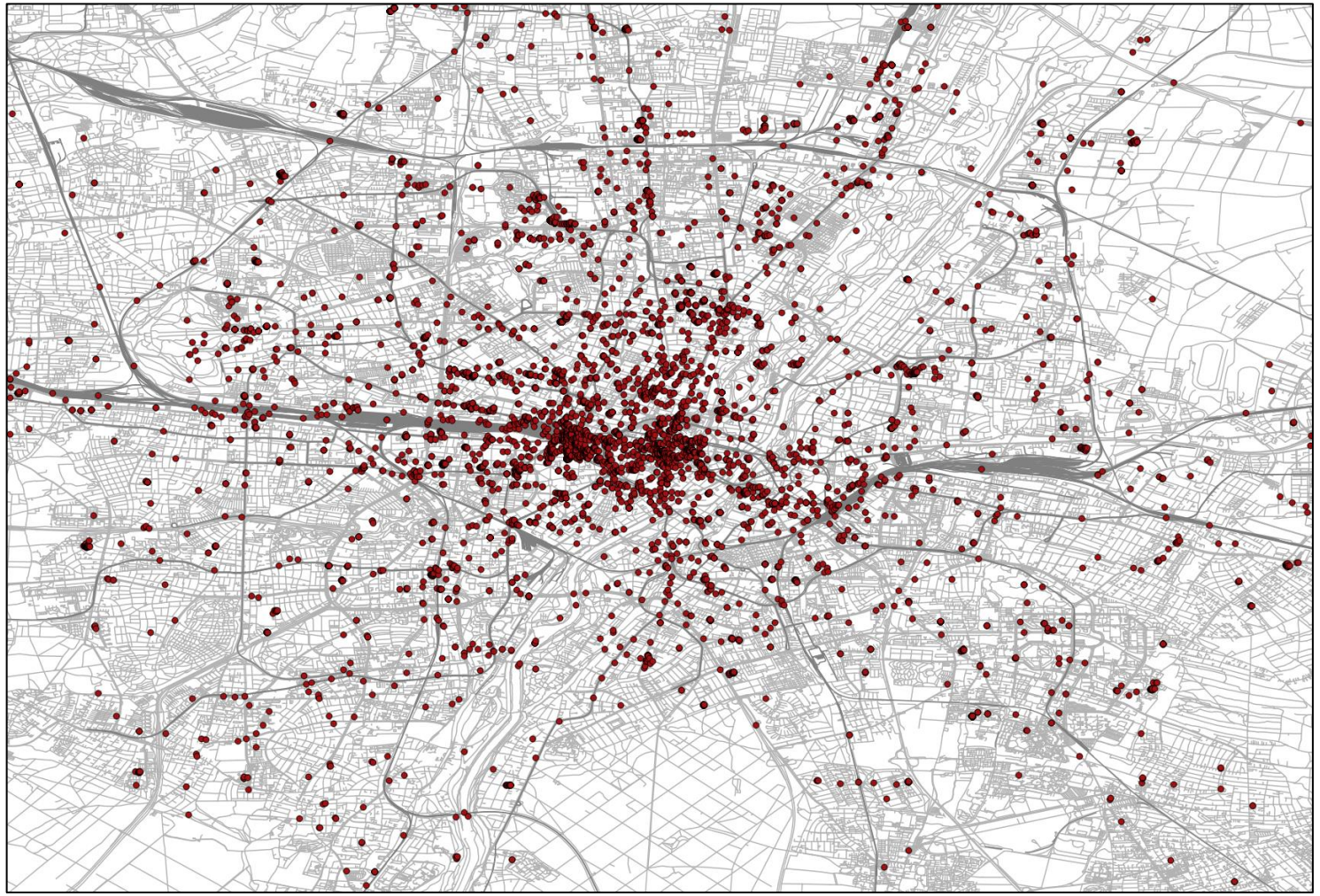
What topics do people discuss?

What are new/trending topics?

Who are the opinion leaders?

How does information spread in the city?

Data Hopes and Dreams

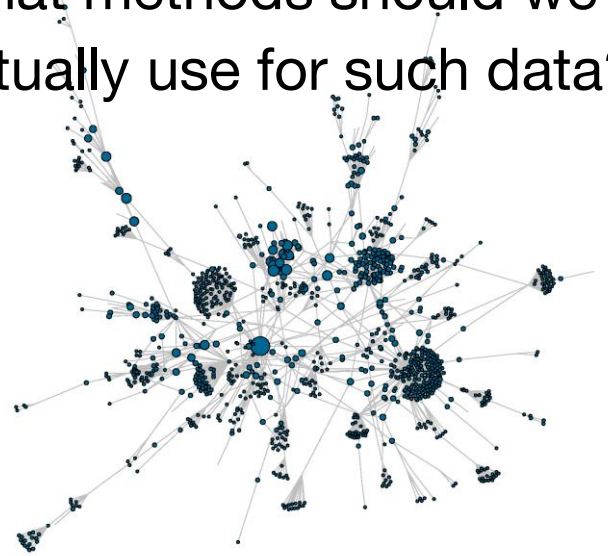


My Actual

Research Questions?

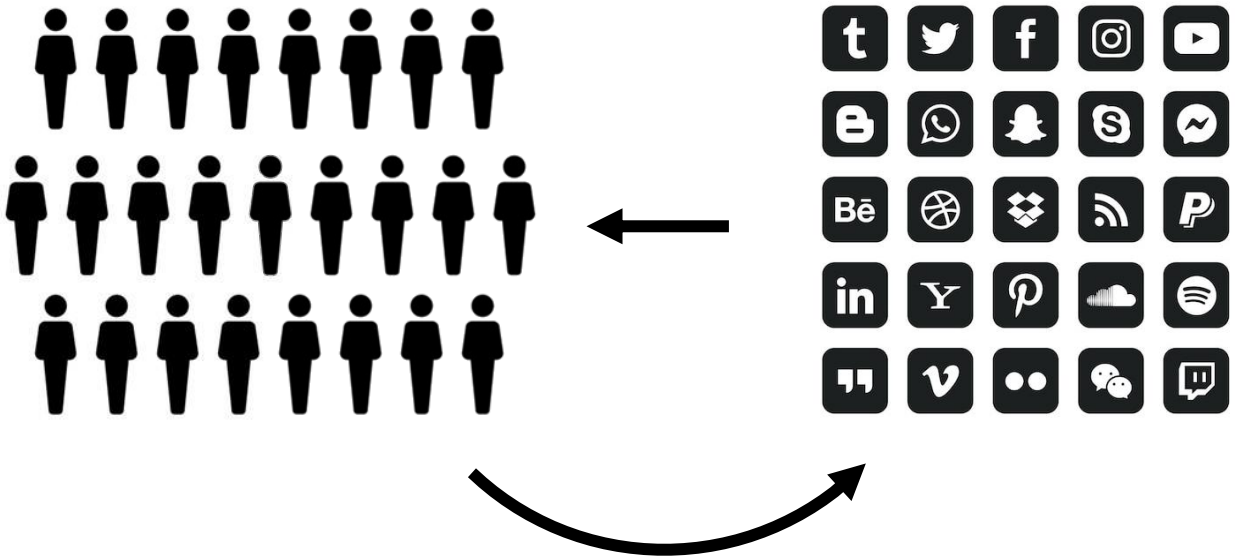
What actually are these little dots?

What methods should we actually use for such data?



Using Social Media for Large-Scale Studies of Human Behavior

The underlying assumption:



Big Data/Social Media Challenge

Representation of Human Populations

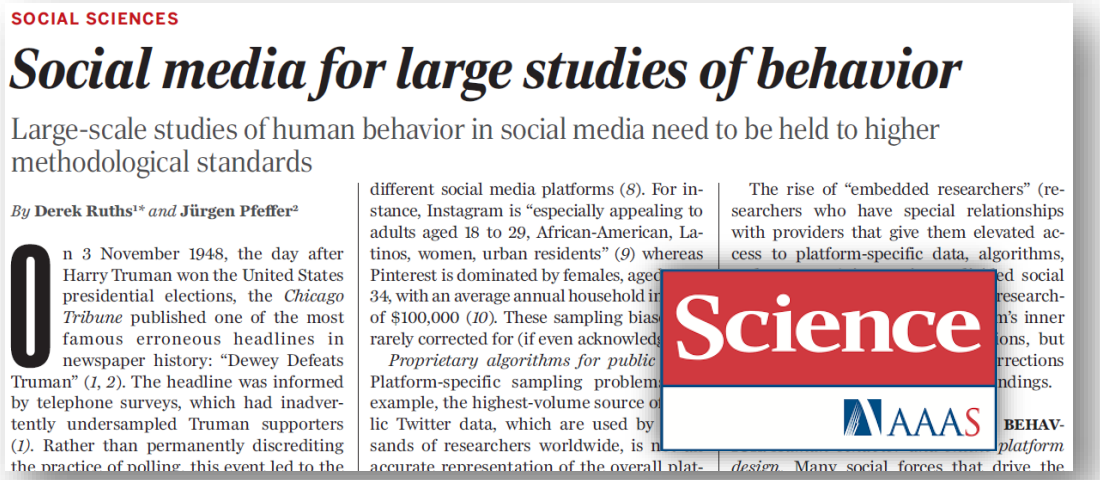
- *Population bias*
- *Proprietary algorithms for public data*

Representation of Human Behavior

- *Human behavior and online platform design*
- *Distortion of human behavior*
- *Non-humans in large-scale studies*

Issues with Methods

- *Proxy population mismatch*
- *Incomparability of methods and data*
- *Multiple comparison problems*
- *Multiple hypothesis testing*



SOCIAL SCIENCES

Social media for large studies of behavior

Large-scale studies of human behavior in social media need to be held to higher methodological standards

By **Derek Ruths**[1]* *and* **Jürgen Pfeffer**[2]

On 3 November 1948, the day after Harry Truman won the United States presidential elections, the *Chicago Tribune* published one of the most famous erroneous headlines in newspaper history: "Dewey Defeats Truman" (*1*, *2*). The headline was informed by telephone surveys, which had inadvertently undersampled Truman supporters (*1*). Rather than permanently discrediting the practice of polling, this event led to the

different social media platforms (*8*). For instance, Instagram is "especially appealing to adults aged 18 to 29, African-American, Latinos, women, urban residents" (*9*) whereas Pinterest is dominated by females, aged 34, with an average annual household income of $100,000 (*10*). These sampling biases are rarely corrected for (if even acknowledged).

Proprietary algorithms for public

Platform-specific sampling problems. For example, the highest-volume source of public Twitter data, which are used by thousands of researchers worldwide, is not an accurate representation of the overall plat-

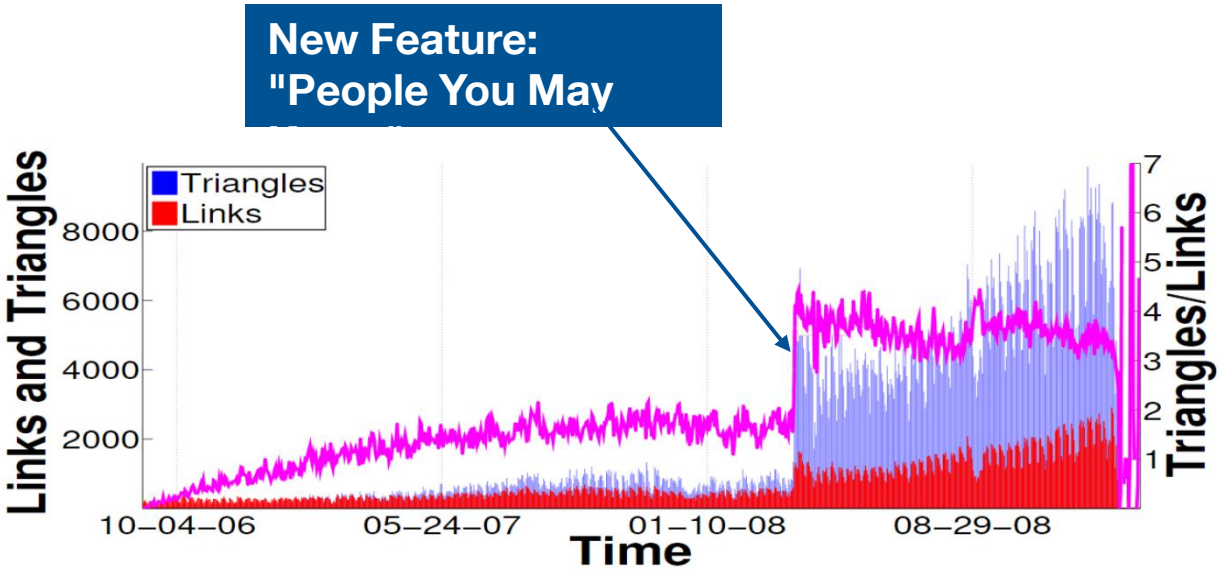
The rise of "embedded researchers" (researchers who have special relationships with providers that give them elevated access to platform-specific data, algorithms,

Science

AAAS

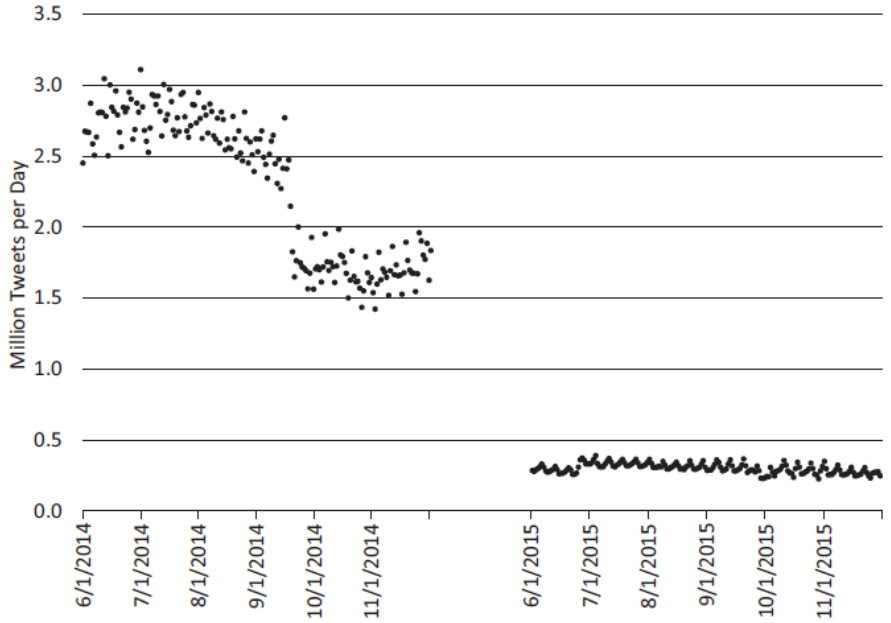
Ruths, Derek & Pfeffer, Juergen (2014). Social Media for Large Studies of Behavior. Science Vol. 346, Issue 6213, 1063 – 1064.

Human Behavior?



Facebook New Orleans

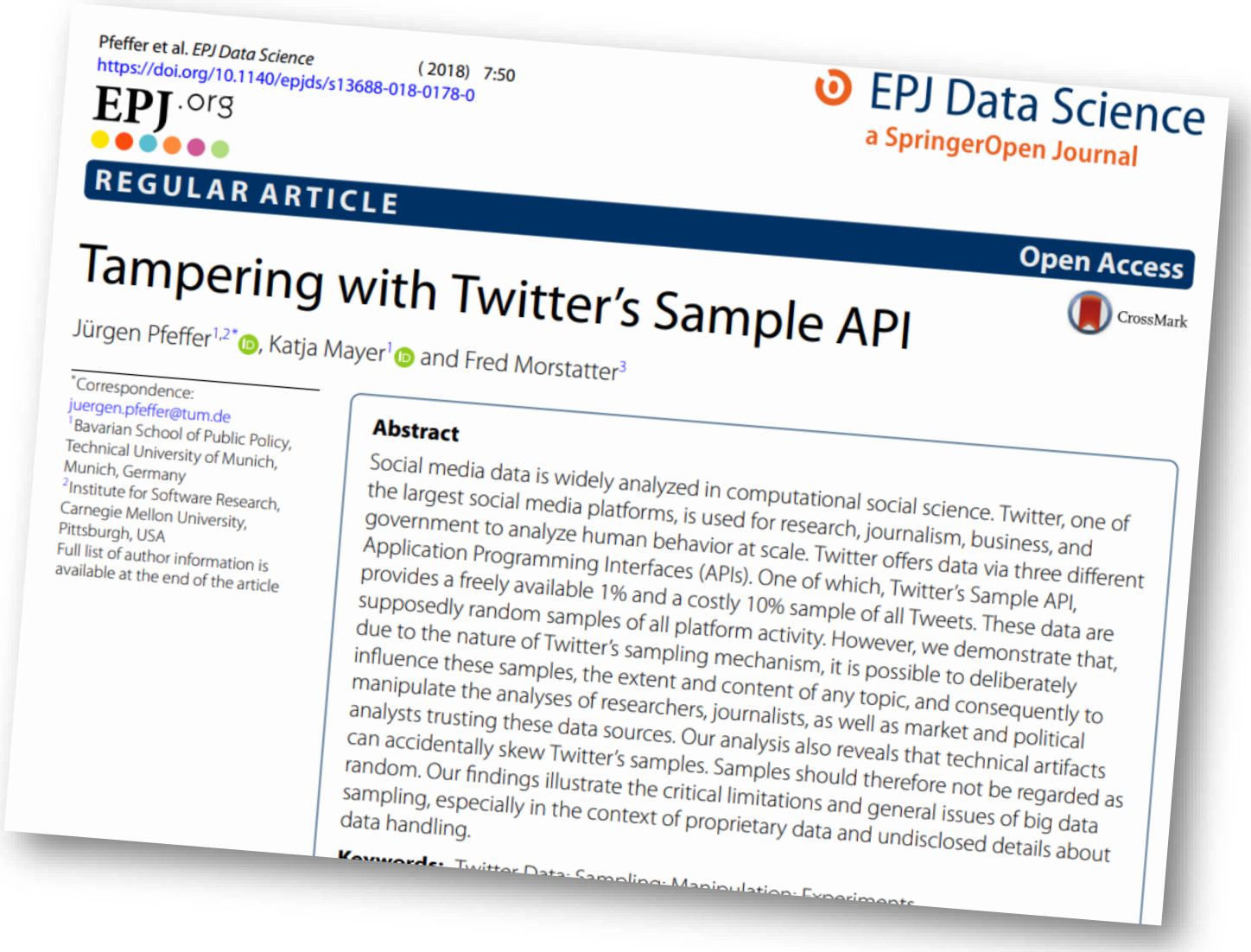
Zignani et al., 2014. Link and triadic closure delay: Temporal metrics for social network dynamics, ICWSM 2014



Geo-Tagged Tweets in the US

Kinder-Kurlanda, Pfeffer, et al. (2017). Archiving information from geotagged tweets to promote reproducibility and comparability in social media research. Big Data & Society, 4(2).

Data Bias from Platform Processes / Sampling



Pfeffer et al. *EPJ Data Science* (2018) 7:50
https://doi.org/10.1140/epjds/s13688-018-0178-0
EPJ.org
EPJ Data Science
a SpringerOpen Journal
REGULAR ARTICLE
Open Access

Tampering with Twitter's Sample API

Jürgen Pfeffer[1,2*], Katja Mayer[1] and Fred Morstatter[3]

*Correspondence:
juergen.pfeffer@tum.de
[1]Bavarian School of Public Policy, Technical University of Munich, Munich, Germany
[2]Institute for Software Research, Carnegie Mellon University, Pittsburgh, USA
Full list of author information is available at the end of the article

Abstract

Social media data is widely analyzed in computational social science. Twitter, one of the largest social media platforms, is used for research, journalism, business, and government to analyze human behavior at scale. Twitter offers data via three different Application Programming Interfaces (APIs). One of which, Twitter's Sample API, provides a freely available 1% and a costly 10% sample of all Tweets. These data are supposedly random samples of all platform activity. However, we demonstrate that, due to the nature of Twitter's sampling mechanism, it is possible to deliberately influence these samples, the extent and content of any topic, and consequently to manipulate the analyses of researchers, journalists, as well as market and political analysts trusting these data sources. Our analysis also reveals that technical artifacts can accidentally skew Twitter's samples. Samples should therefore not be regarded as random. Our findings illustrate the critical limitations and general issues of big data sampling, especially in the context of proprietary data and undisclosed details about data handling.

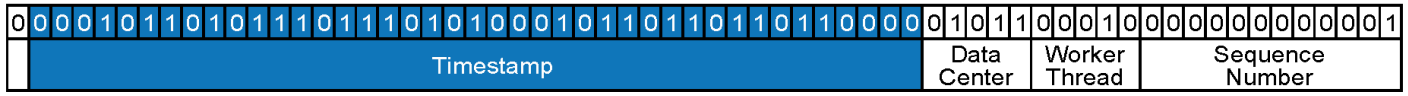
How does Twitter's Sample API's Sampling Work?

Tweet:
„Thank you for everything. My last ask is the same as my first. I'm asking you to believe—
not in my ability to create change, but in yours.“
@POTUS44, Jan 11, 2017

TweetID: 819044196371800065

↓

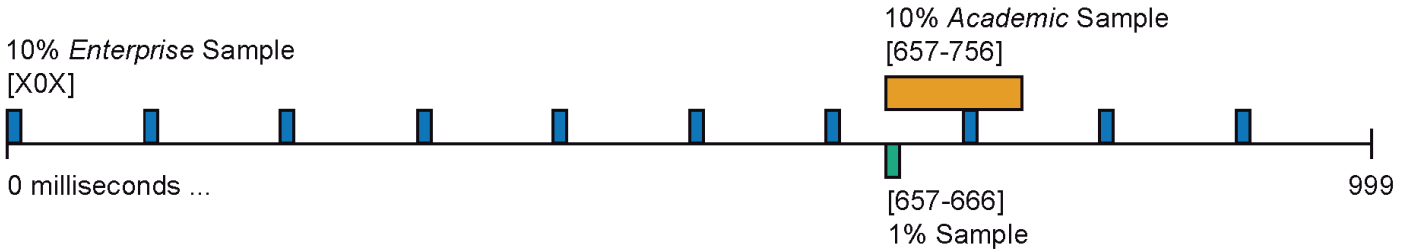
64 bit representation of Tweet ID:



Integer representation timestamp: 1484110328177
Milliseconds of timestamp: 177

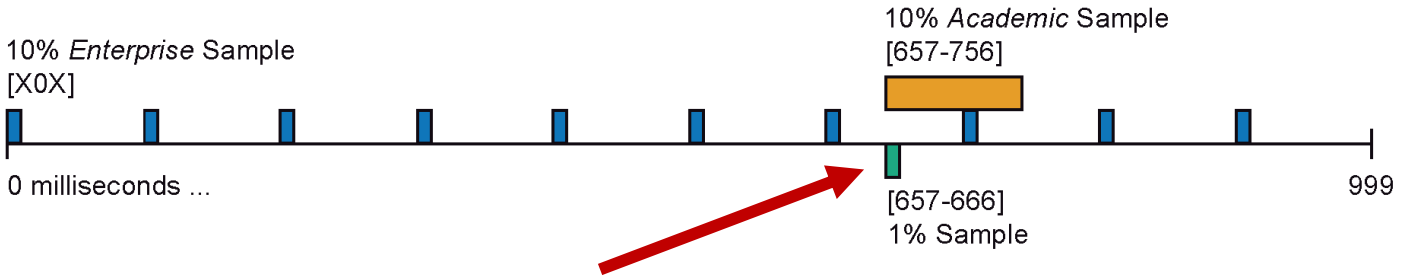
↓

Sampling based on millisecond selection windows:



Kergl, D. et al. (ASONAM 2014), Morstatter, F., et al. (WWW 2016)

Can we Manipulate the Sample?



Can we hit the 10ms selection window at high rate?

100 Twitter accounts

Limited resources

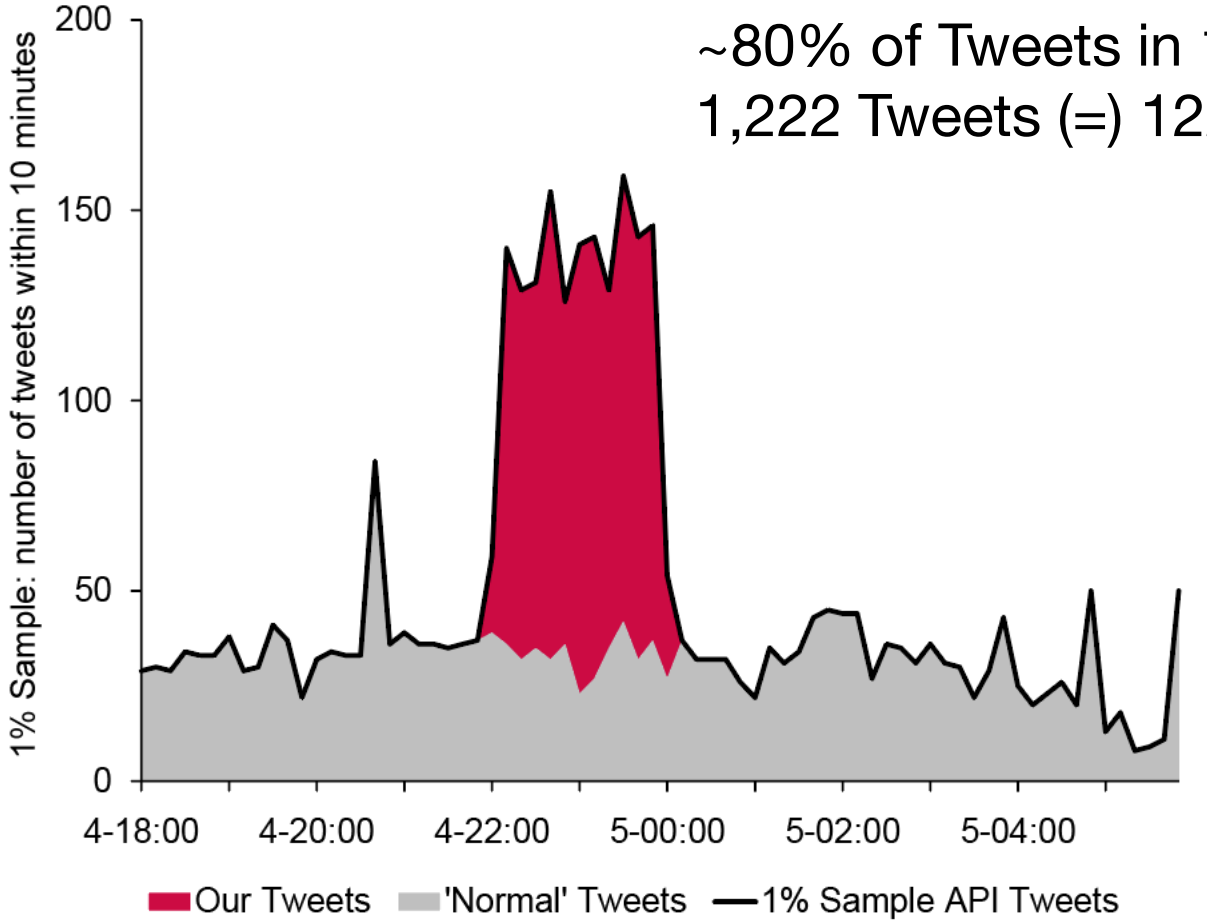
Limit impact (ethics statement, info sharing)

known from API

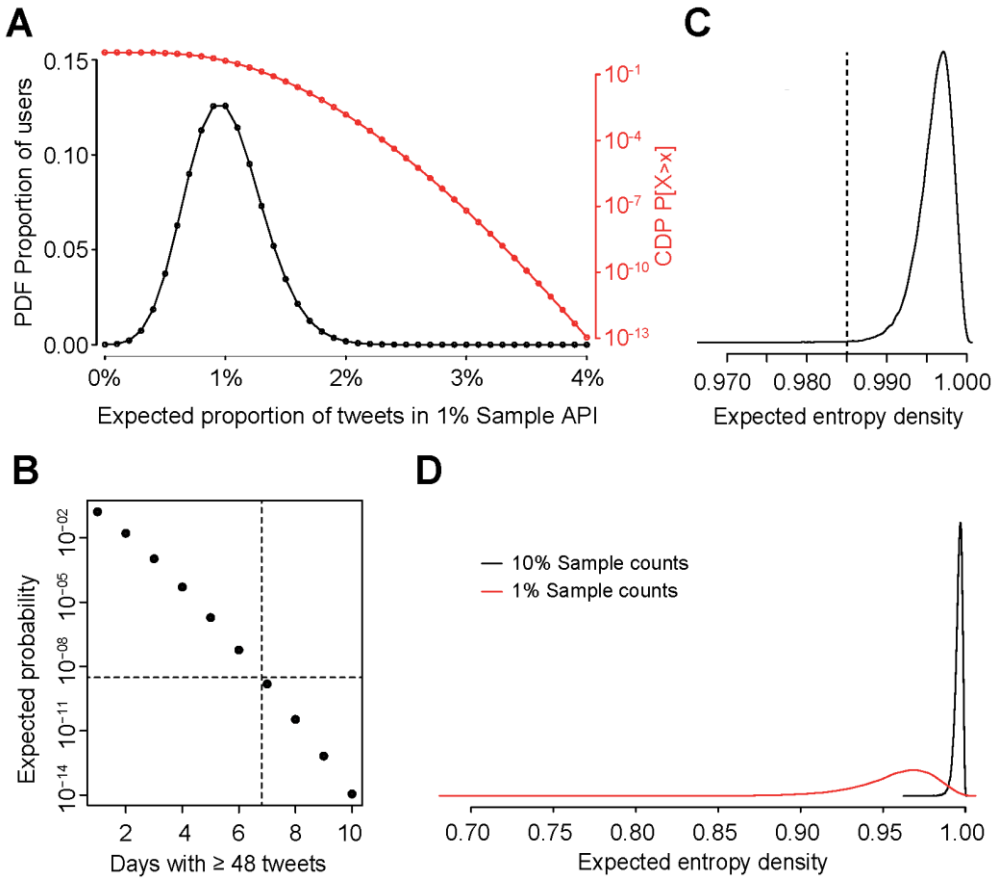
known locally

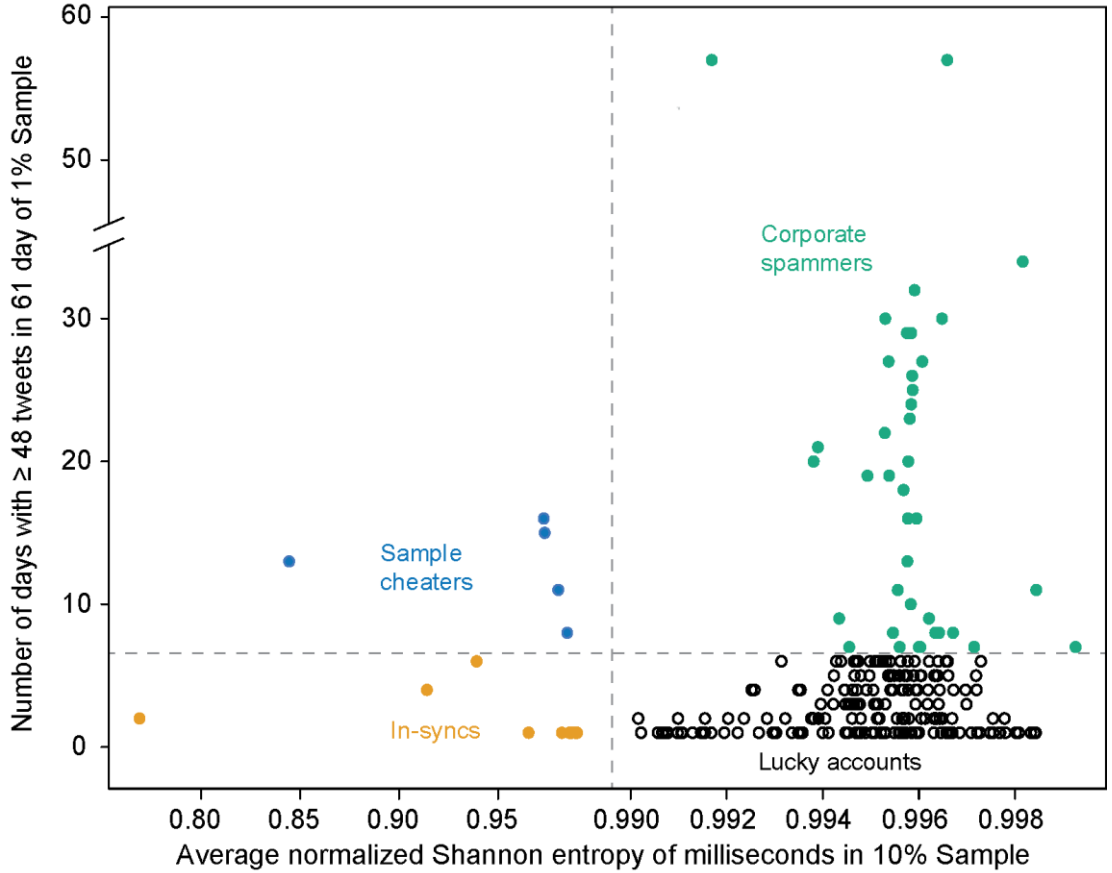
$$t_a = t_s + t_\delta$$

Tampering Experiments



Identifying Over-Represented Accounts





In and Out of Sync

Imagine… a lab Twitter bot

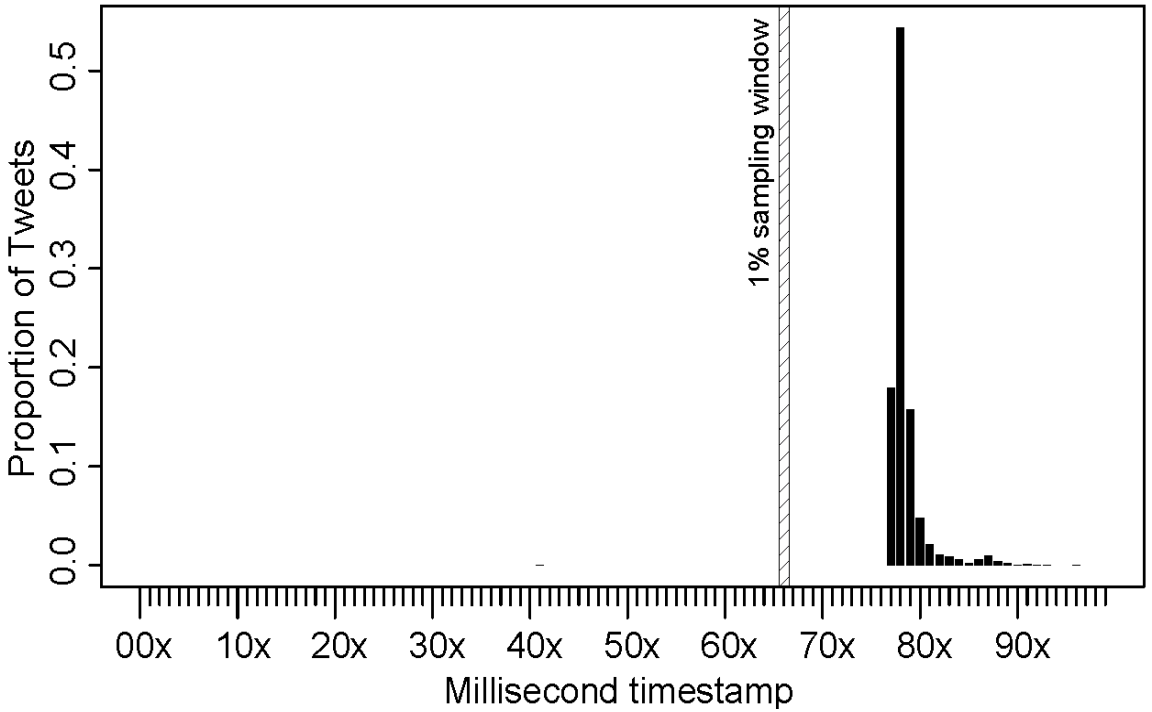
Tweeting the lab temperature

Exactly every 5 minutes

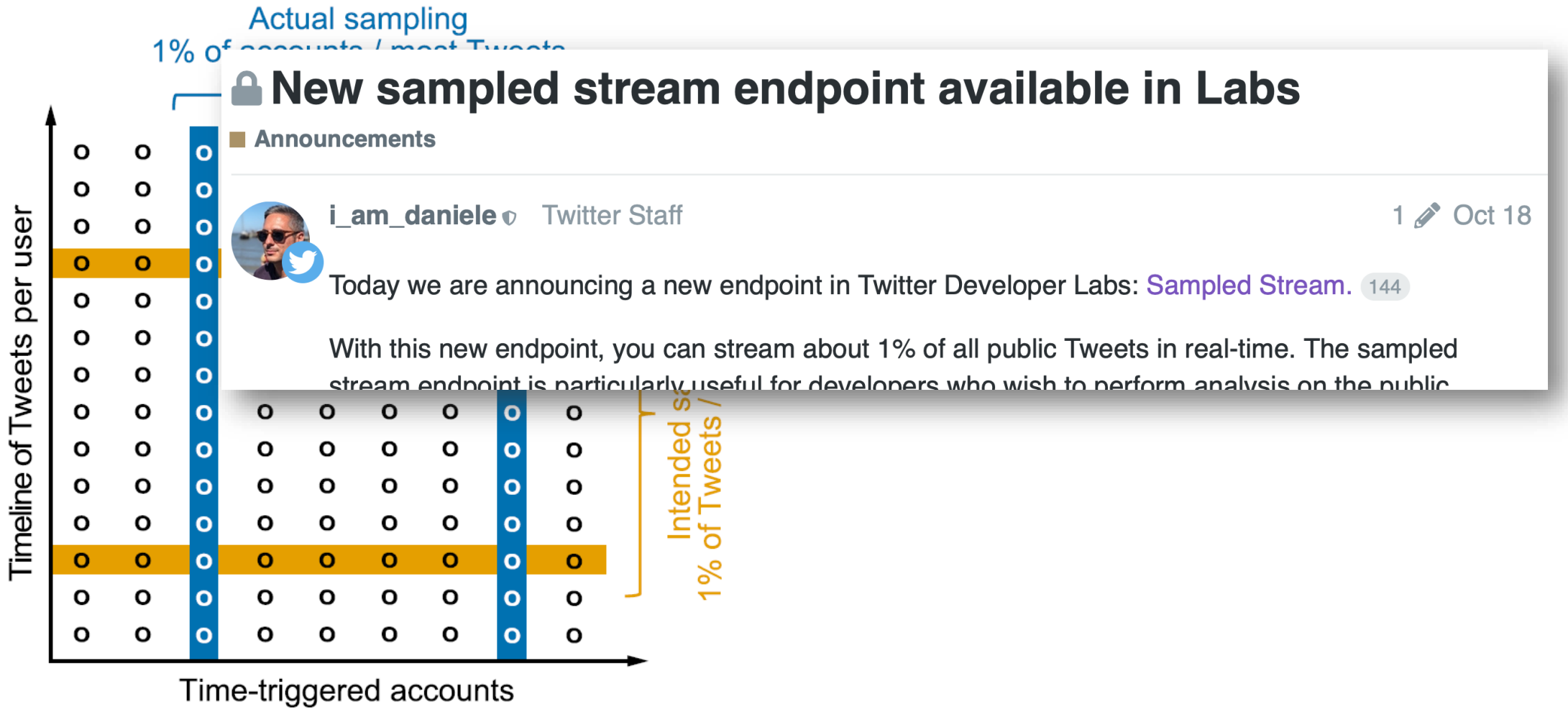
10,000 times

How many Tweets will be in 1% Sample?

Answer: Most likely 0



The Flipped Sample



Conclusions I: Bias by Design & Bias by Purpose

Algorithms inherit the biases of their creators

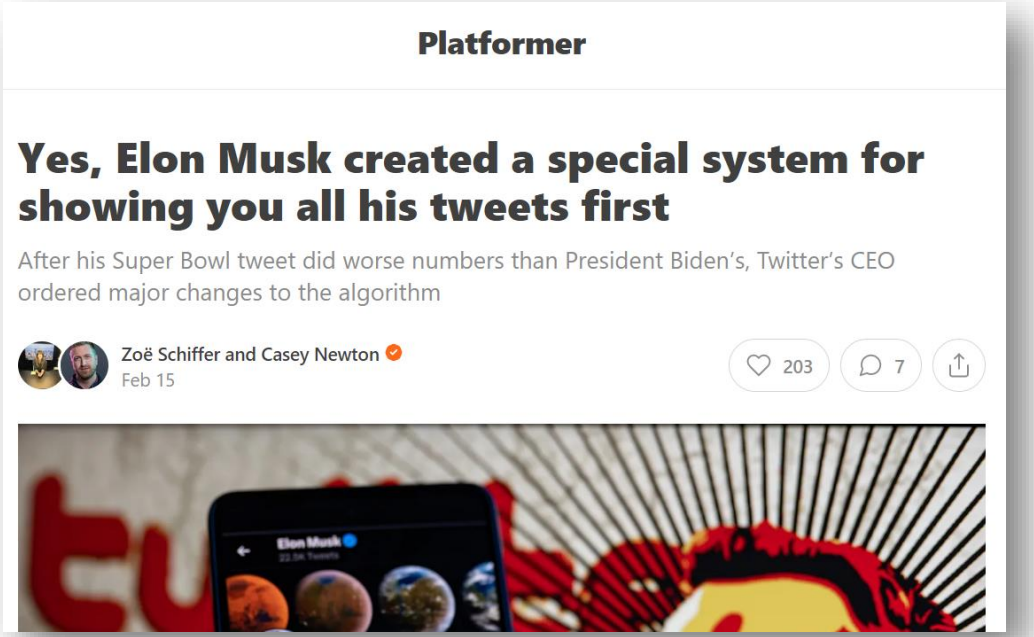
Bias from unintentional side effects of systems

Critical issues that come from…

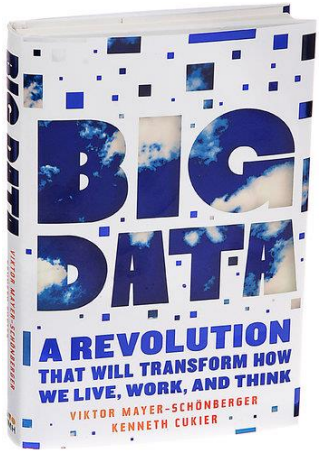
- Working with proprietary data
- Undisclosed details about data handling

“People ain’t no good.” *(Nick Cave, 1997)*

Behavior of many or manipulation of a few?



Can we Ever Know the Data?



There is no "Full" dataset of human behavior

Many un-known technical layers

Samples:
Not **"an artifact of a period of information scarcity"**

N = All

"The concept of sampling no longer makes as much sense when we can harness large amounts of data."

(Schönberger & Cukier, 2013, p.26)

N = All. Is it all?
All of what?
Is it all of what we want?
Is it all of what we think it is?
Is it anything that we are actually interested in?

Overcoming Sampling Issu

Just Another Day on Twitter: A Complete 24 Hours

Jürgen Pfeffer[1], Daniel Matter[1], Kokil Jaidka[2], Onur Varol[3], Afra M
Dennis Assenmacher[6], Siqi Wu[7], Diyi Yang[8], Cornelia Brantner[9], D
Otterbacher[10], Carsten Schwemmer[11], Kenneth Joseph[12], David G
[1]School of Social Sciences and Technology, Technical University of Munich, [2]Centre f
National University of Singapore, [3]Sabanci University, [4]University of Washington (Both
[6]GESIS - Leibniz Institute for the Social Sciences, [7]University of Michigan, [8]Stanfor
[10]Open University of Cyprus & CYENS CoE, [11]Ludwig Maximilian University of
[13]University of Konstanz, [14]Information Sciences Institute, University of Southern Ca
Vienna

Abstract

At the end of October 2022, Elon Musk concluded his acquisition of Twitter. In the weeks and months before that, several questions were publicly discussed that were not only of interest to the platform's future buyers, but also of high relevance to the Computational Social Science research community. For example, how many active users does the platform have? What percentage of accounts on the site are bots? And, what are the dominating topics and sub-topical spheres on the platform? In a globally coordinated effort of 80 scholars to shed light on these questions, and to offer a dataset that will equip other researchers to do the same, we have collected *all*

AAAI International
(ICWSM), in the pa
30 scientific papers
range of topics rangi
yses to politics and
gence, Twitter has
(i.e., social telescop
ence (Mejova, Web
for social science th
global patterns of c
hourly changes in e

Table 2: The categories of the top 500 hashtags in the dataset

Category	Hashtags	Occurrence	
Celebrities	159	20,809,742	25.5%
Sex	104	20,529,196	25.2%
Iranian Protests	15	13,488,295	16.6%
Entertainment	45	4,392,227	5.4%
Advertisement	32	4,644,540	5.7%
Politics	38	3,858,550	4.7%
Finance	30	3,549,107	4.4%
Games	21	3,348,128	4.1%
Other	31	2,672,291	3.3%
Unknown	25	4,176,432	5.1%
Sum	500	81,468,508	100.0%

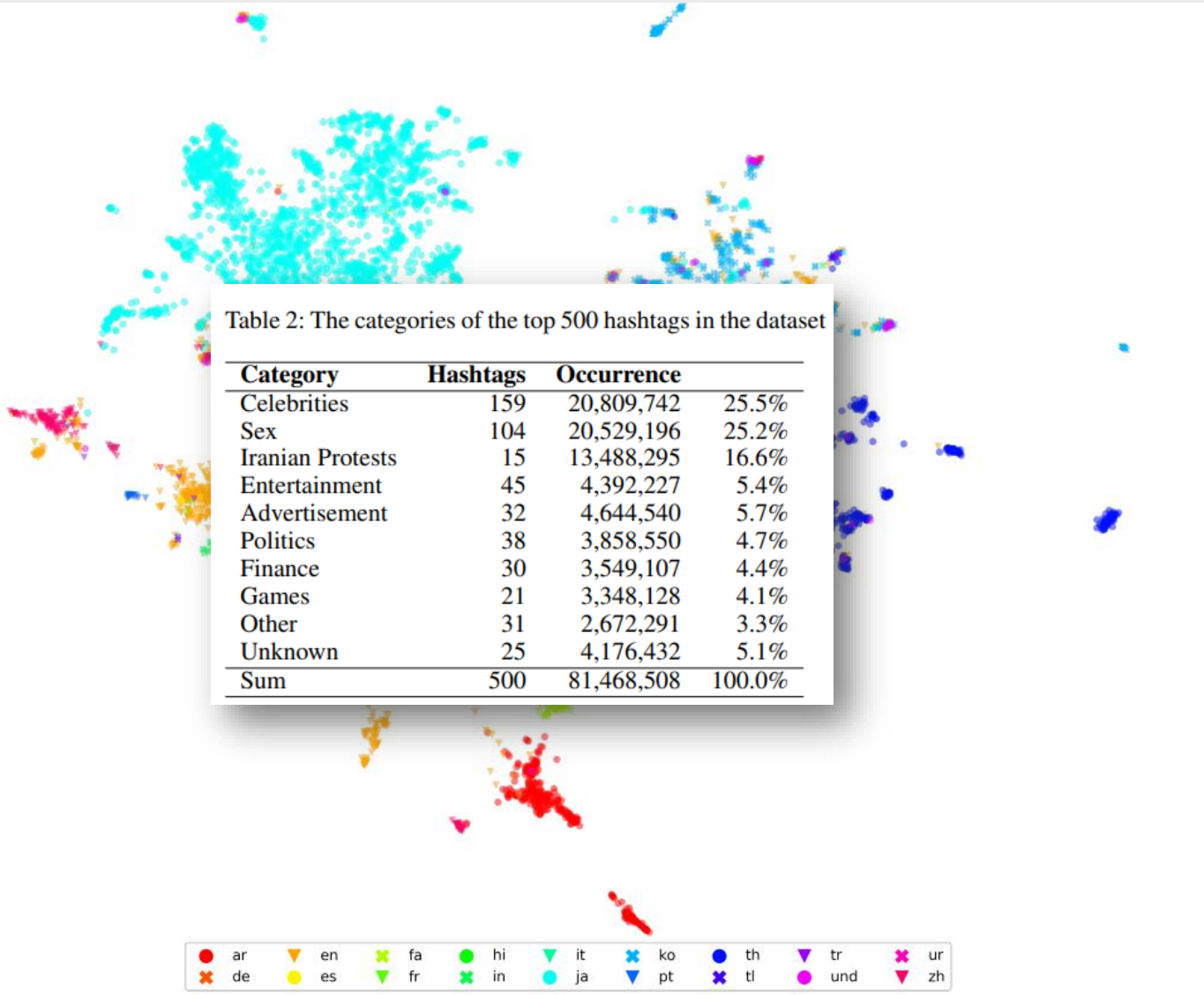


Figure 5: MDS of top 10,000 hashtags based on co-usage by same accounts; colors represent dominant language in tweets using a hashtag.

Data Collection in a Post-API World

April 2015: Facebook closure of Public Search API (all public posts, 2 weeks back)

April 2018: Facebook closure of Pages API (public pages)

February 2023 announcement: Twitter will close APIs

New TikTok Research API: Restricted, controlled

New Facebook Research API: ?

Scraping!

So, where do we go from here?

Unresolved issues: Black-boxed systems, hidden algorithms, sampling, biases, etc.

Post API world: Access limitations, less standards, less reproducibility, more un-knowns

How can we study human behavior with data in the future?

Collaborate with Smaller Platforms

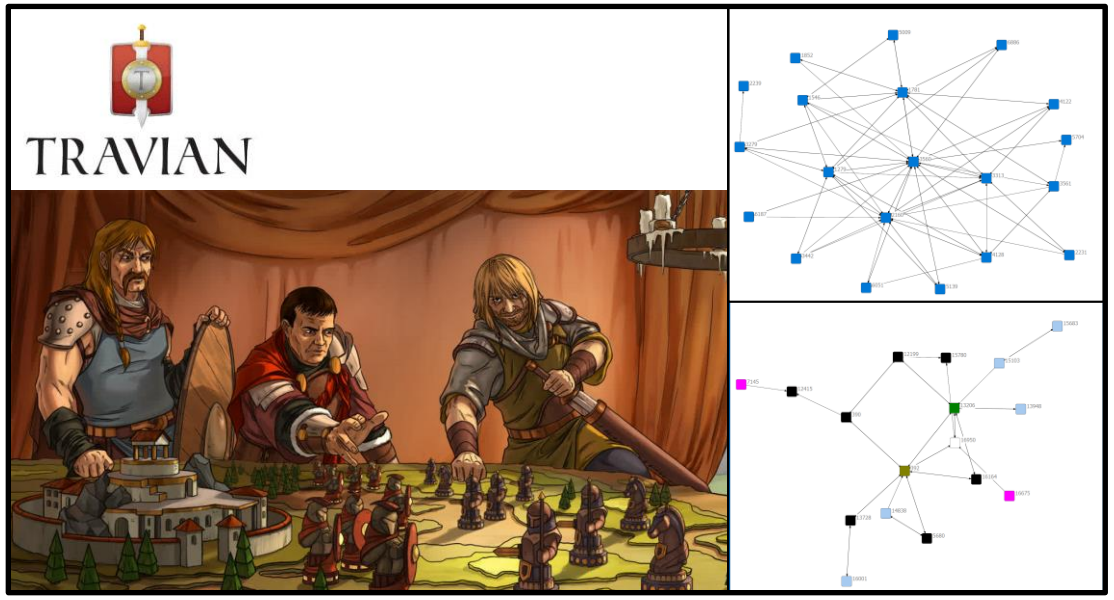
User Reported Drug Experiences

Platform for documenting experience with psychoactive substances: erowid.org
Number of reviewed drug reports: 36,713

Mooseder, A., Malik, M.M., Lamba, H., Erowid, E., Thyssen, S., Pfeffer, J., (2022). Glowing Experience or Bad Trip? A Quantitative Analysis of User Reported Drug Experiences on Erowid.org. ICWSM 2022, pp. 675-686.

Leadership & Success in MMOG Teams

Collaborative browser-based strategy game Travian
Game duration: 1 year
Number of teams in our data: 4,758 (avg. size 14.5)



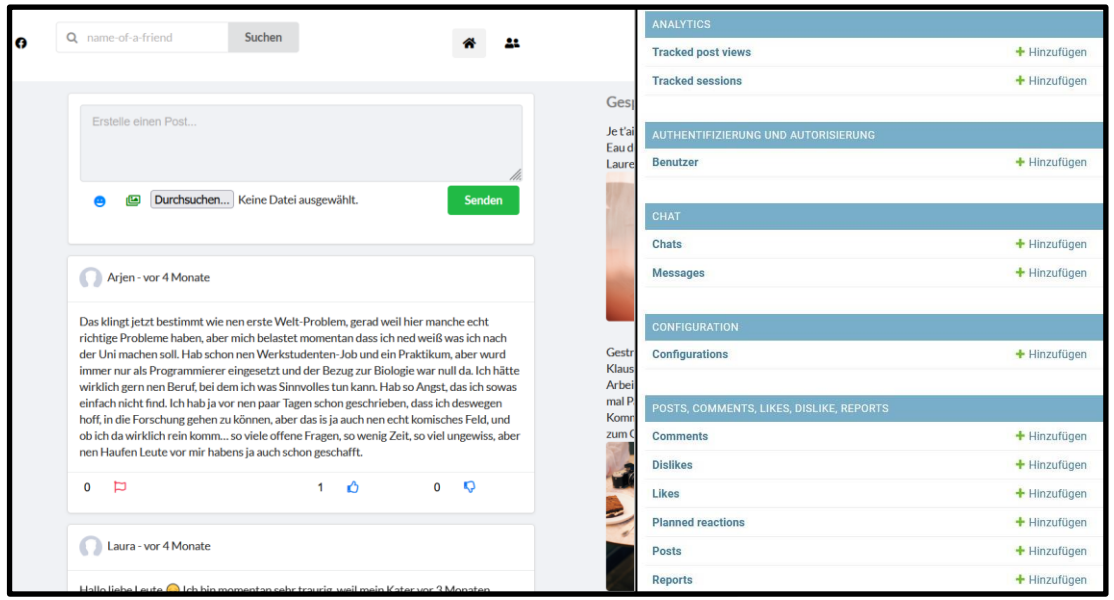
Müller S, Ghawi R, Pfeffer J (2023) Reviewing the potentials of MMOGs as research environments: A case study from the strategy game Travian. PLoS ONE 18(2): e0281114.

Clone Platforms for Experiments

Fakebook – A Facebook for Research

Look & Feel of Facebook, but researcher can:

- Track all user data (posts, likes, views, …)
- Control the content (e.g., create adds)
- Control the platform (e.g., what users can see)
- Create planned actions (posts and likes which are executed by bots at a specified time)



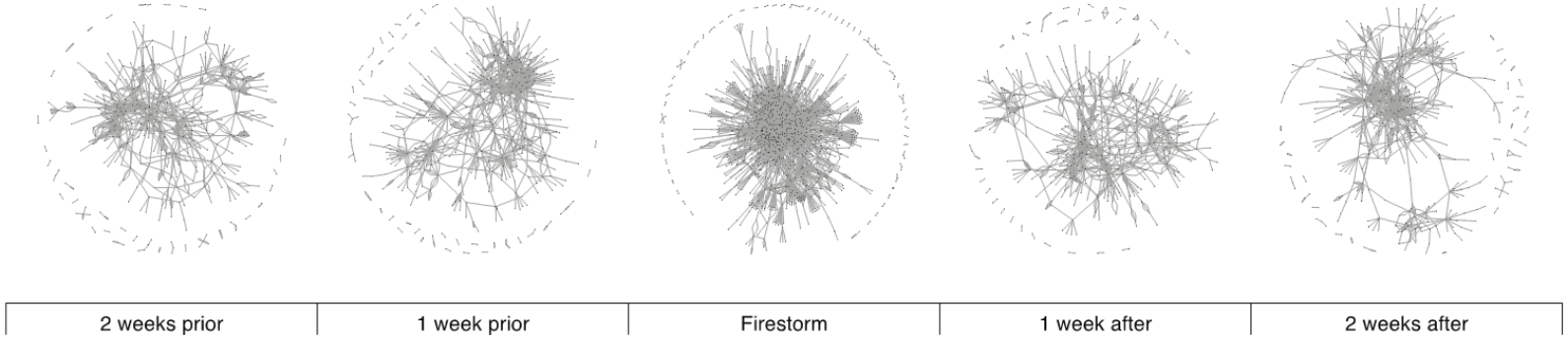
Current study:
The offline impact of online feedback

"People receiving no likes had significantly less fun and felt more stressed and sad during the interaction on Fakebook."

→ Make tools available for research.

Dynamics: Network Change Analysis

Online Firestorm on Twitter



Post Merger Integration in Real Companies

Two organizations in the chip and semiconductor industry
Six month of data
3,807 employees, 21 attributes (age, education, position, …)
250,455 emails
Study effects of post-merger intra-company communication

How does change happen?
How can we model and analyze change processes?

Co-Creation Processes - Online Misogyny

Summary

Good data barely comes off-the-shelf

Don't be satisfied with black-boxed secondary data

It is often more than data: A complex socio-technical system

Opportunities of a possible post-API world:

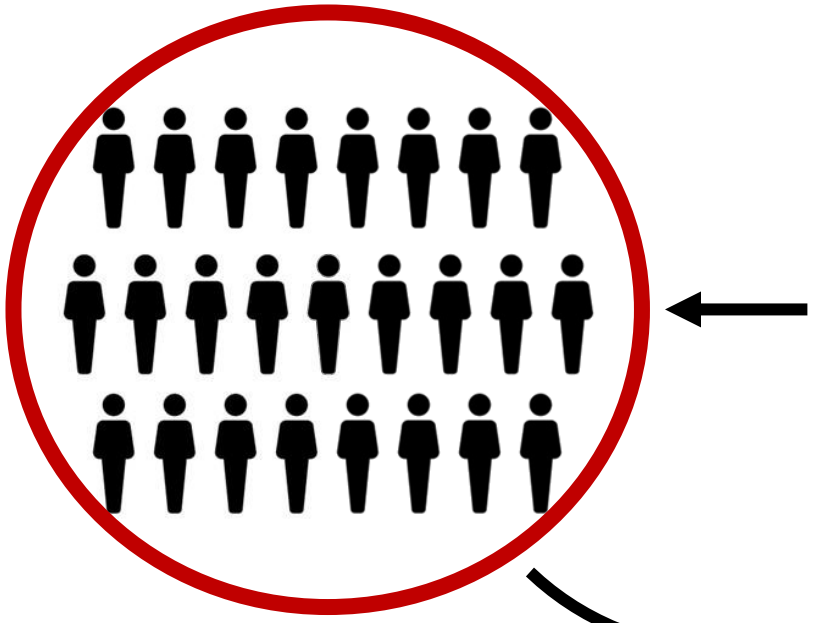
- Less: "Collect first, think later“; less (garbage) data
- Which data do we really need for our research questions?
- New data sources, new approaches

Outlook

Next step: Get closer to the data

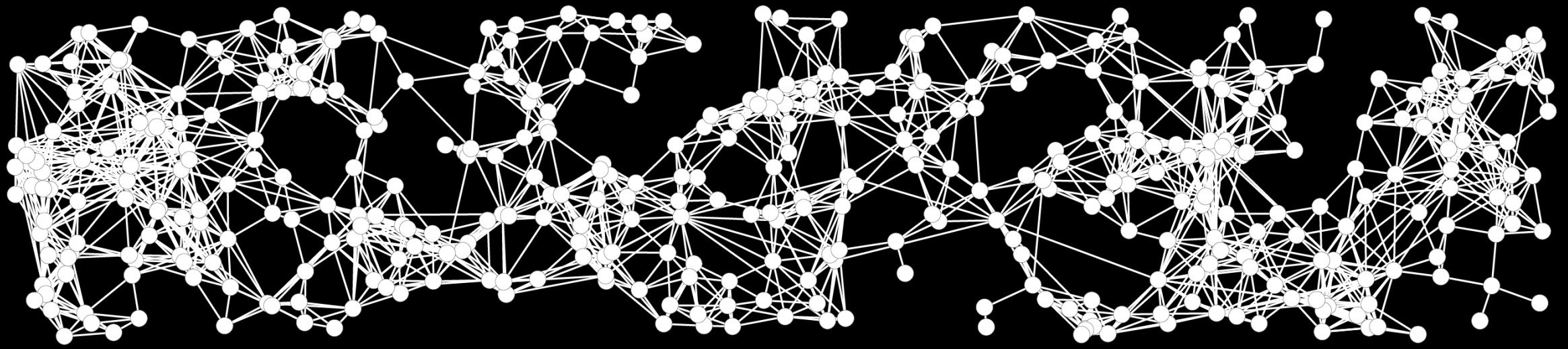
- Collaborate with people creating and handling data
- Create “primary” data with own tools
- Co-creation processes with stakeholders

Studying Human Behavior with Data



We don’t do data science, we do people science with data!

“Our mission is to go forward, and it has only just begun.
There's still much to do, still so much to learn. Engage!”
Jean-Luc Picard, Star Trek TNG, Season 1 Episode 26



Jürgen Pfeffer
Juergen.Pfeffer@tum.de
@JurgenPfeffer