

Quantitative Text Analysis. Applications to Social Media Research

Pablo Barberá

London School of Economics

www.pablobarbera.com

Course website:

pablobarbera.com/text-analysis-vienna







George Takei

March 28 at 10:10pm ·

Who's with me.



Like · Comment · Share

408,735 people like this.

66,990 shares



Bon Alimago
@karma_thief

I need a hug. I have never been so traumatized by a television show.
#gameofthrones

More

RETWEETS 356 FAVORITES 110



10:06 PM - 2 Jun 2013



Google

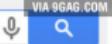
how do I convert to

how do I convert to judaism

how do I convert to islam

how do I convert to catholicism

how do I convert to pdf



VIA 9GAG.COM

Press Enter to search.



Justin Bieber
@justinbieber

I make music. I love music.

More

RETWEETS 54,213 FAVORITES 59,205



10:09 PM - 7 Apr 2014



dustin curtis

@dcurtis



Follow

"At any moment, Justin Bieber uses 3% of our infrastructure. Racks of servers are dedicated to him. - A guy who works at Twitter

RETWEETS

1,528

FAVORITES

267



8:56 PM - 6 Sep 2010



...



Dmitry Medvedev @MedvedevRussiaE

Follow

The harmonious development of Crimea and Sevastopol as part of our state is one of the main objectives of the Russian Government

Reply Retweet Favorite More

RETWEETS
144

FAVORITES
57



10:39 AM - 21 Mar 2014



The New York Times
April 2

"Much of the foreign media coverage has distorted the reality of my country and the facts surrounding the events," writes Nicolás Maduro, the president of Venezuela, in Opinion: <http://nyti.ms/1gP5o2I>

Like · Comment · Share

57

262 people like this.

Top Comments ▾



Elizabeth Warren shared a link.
January 16

I'm not giving up on our fight to extend unemployment benefits. Watch my interview with Now With Alex Wagner about why we need to keep fighting.



Warren: This is the moment to back on economy
www.msnbc.com

President Obama faces one huge problem with his effort to improve the economy: an opposition party

Like · Comment · Share

15,483 720 1,041



Jackie Walorski @RepWalorski

Follow

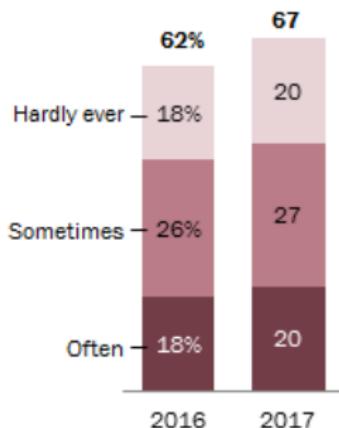
Today, a representative from my office will be meeting with constituents in Goshen. For more details, visit walorski.house.gov/services/upcom...

Reply Retweet Favorite More

11:22 AM - 8 Apr 2014

In 2017, two-thirds of U.S. adults get news from social media

% of U.S. adults who get news from social media sites ...



Source: Survey conducted Aug. 8-21, 2017.
"News Use Across Social Media Platforms 2017"

- ▶ 67% of Americans get news on social media (Pew Research)
- ▶ 58% of EU citizens active on social media & find it useful to get news on national political matters (Eurobarometer, Fall 2017)
- ▶ Social media: top source of news for U.S. young adults (Pew)



Shift in communication patterns



Digital footprints of human behavior

This course

Two central questions:

1. **What** type of social science questions can I answer with social media text? (**Today**)
2. **How** would I answer those questions? What methods and tools would I use?

Tomorrow

- ▶ Introduction to text analysis with R. Dictionary methods

Wednesday

- ▶ Large-scale text classification with supervised (machine learning) and unsupervised (topic models) methods

Thursday

- ▶ Collecting social media data with R (Twitter)

Friday

- ▶ Advanced topics in text analysis

About me: Pablo Barberá

- ▶ Assistant Professor of Computational Social Science at the [London School of Economics](#)
 - ▶ Previously Assistant Prof. at [Univ. of Southern California](#)
 - ▶ PhD in Politics, [New York University](#) (2015)
 - ▶ Data Science Fellow at [NYU](#), 2015–2016
- ▶ [My research:](#)
 - ▶ Social media and politics, comparative electoral behavior
 - ▶ Text as data methods, social network analysis, Bayesian statistics
 - ▶ Author of R packages to analyze data from social media
- ▶ [Contact:](#)
 - ▶ P.Barbera@lse.ac.uk
 - ▶ www.pablobarbera.com
 - ▶ [@p_barbera](https://twitter.com/p_barbera)

Your turn!



1. Name?
2. Affiliation? Background?
3. Summarize your research interests in 5 words

Course philosophy

How to learn the techniques in this course?

- ▶ Lecture approach: not ideal for learning how to code
- ▶ You can only **learn by doing**.
- We will cover each concept three times during each session
 1. Introduction to the topic
 2. Guided coding session
 3. Coding challenges
 - Repeat 2-3 times per day
- ▶ Warning! We will **move fast**.

Course logistics

Evaluation:

- ▶ Attendance and participation: 30%
- ▶ Final paper: 70%
 - ▶ Due by TBC
 - ▶ Goal: analysis of social media text
 - ▶ Length/requirements TBC
 - ▶ Graded on a 100-point scale

Course website

pablobarbera.com/text-analysis-vienna

Social Media Research: Opportunities and Challenges

Social media data

What are the main advantages of using social media data to study human behavior?

1. Unobtrusive data collection at scale, e.g. in study of networks, censorship
2. Homogeneity in data format across actors, countries, and over time, e.g. in study of political rhetoric
3. Temporal and spatial data granularity, e.g. in study of geographic segregation
4. Increasing representativeness of social media users, e.g. in study of political elites

Social media research

Two different approaches in the growing field of social media research:

1. Social media as a new source of data
 - ▶ Behavior, opinions, and latent traits
 - ▶ Interpersonal networks
 - ▶ Elite behavior
 - ▶ Affordable field experiments
2. How social media affects social behavior
 - ▶ Collective action and social movements
 - ▶ Political campaigns
 - ▶ Social capital and interpersonal communication
 - ▶ Political attitudes and behavior

Social media research

Two different approaches in the growing field of social media research:

1. Social media as a new source of data
 - ▶ Behavior, opinions, and latent traits
 - ▶ Interpersonal networks
 - ▶ Elite behavior
 - ▶ Affordable field experiments
2. How social media affects social behavior
 - ▶ Collective action and social movements
 - ▶ Political campaigns
 - ▶ Social capital and interpersonal communication
 - ▶ Political attitudes and behavior

Behavior, opinions, and latent traits

- ▶ Digital footprints: check-ins, conversations, geolocated pictures, likes, shares, retweets, ...
- Non-intrusive measurement of behavior and public opinion



Regular Article

Every tweet counts? How sentiment analysis of social media can improve our knowledge of citizens' political preferences with an application to Italy and France

new media & society
2014, Vol. 16(2) 340–358

© The Author(s) 2013

Reprints and permissions:

sagepub.co.uk/journalsPermissions.nav

DOI: [10.1177/146144813480466](https://doi.org/10.1177/146144813480466)

nms.sagepub.com



Andrea Ceron, Luigi Curini, Stefano M Iacus

Università degli Studi di Milano, Italy

Giuseppe Porro

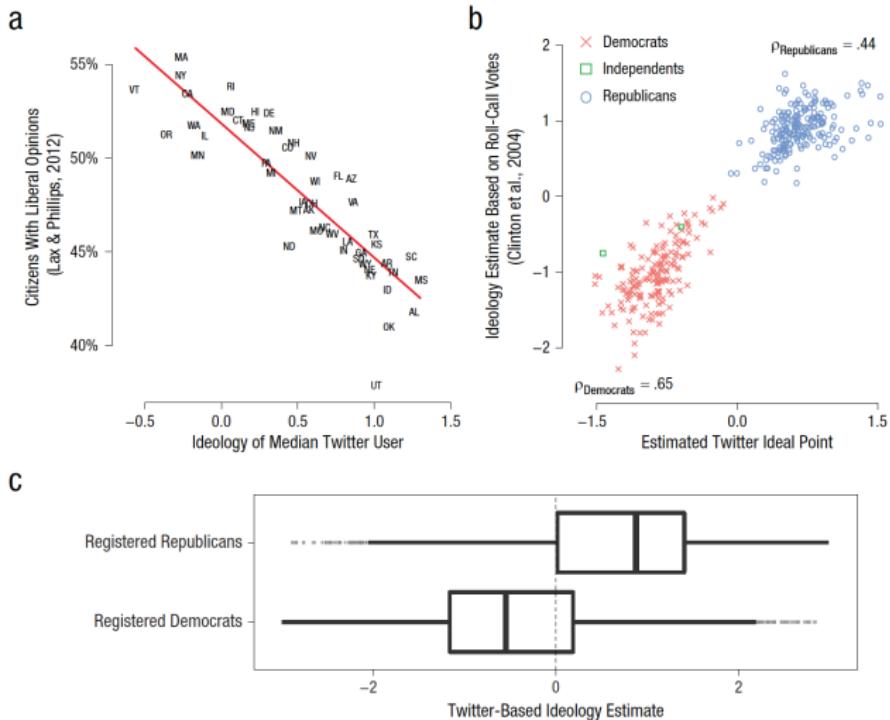
Università degli Studi dell'Insubria, Italy



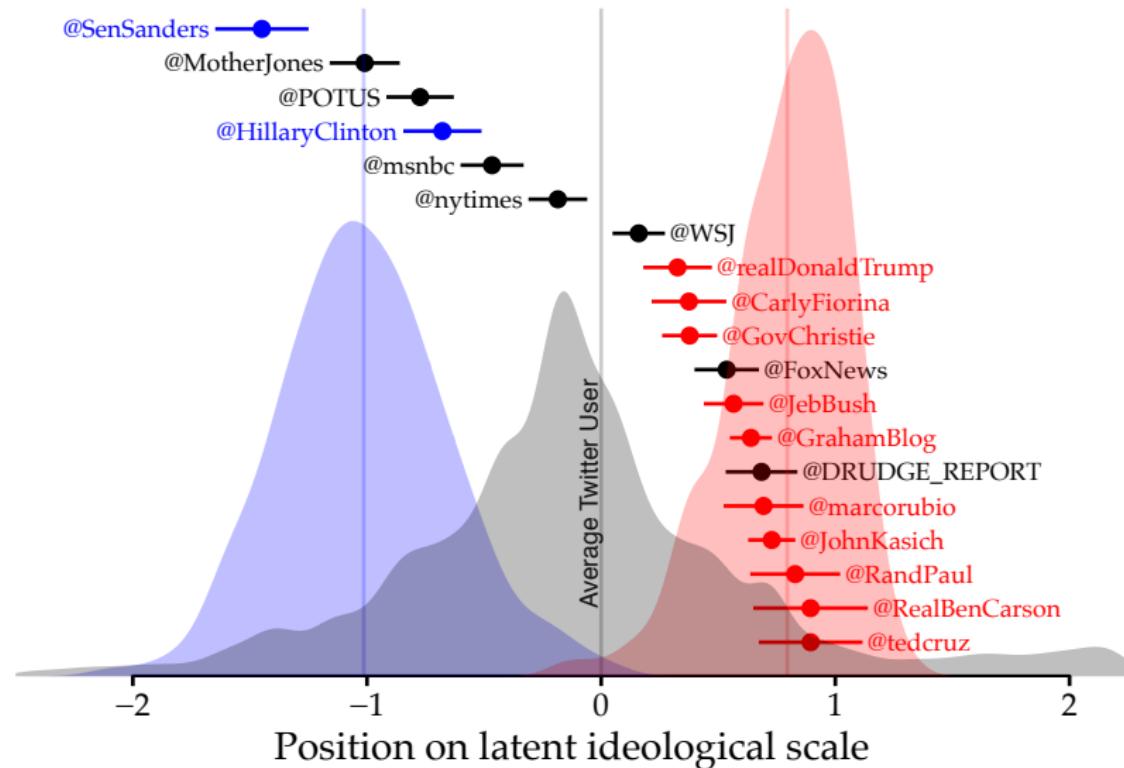
AMERICAN JOURNAL
of POLITICAL SCIENCE

Behavior, opinions, and latent traits

- Inference of latent traits: political knowledge, ideology, personal traits, socially undesirable behavior, . . .



Estimating political ideology using Twitter networks



Barberá “Who is the most conservative Republican candidate for president?” *The Monkey Cage / The Washington Post*, June 16 2015

Social media research

Two different approaches in the growing field of social media research:

1. Social media as a new source of data
 - ▶ Behavior, opinions, and latent traits
 - ▶ **Interpersonal networks**
 - ▶ Elite behavior
 - ▶ Affordable field experiments
2. How social media affects social behavior
 - ▶ Collective action and social movements
 - ▶ Political campaigns
 - ▶ Social capital and interpersonal communication
 - ▶ Political attitudes and behavior

Interpersonal networks

- ▶ Political behavior is social, strongly influenced by peers

Today is Election Day [What's this?](#) • [close](#)

 Find your polling place on the U.S. Politics Page and click the "I Voted" button to tell your friends you voted.

I Voted


0 1 1 5 5 3 7 6
People on Facebook Voted

 **f** Jaime Settle, Jason Jones, and 18 other friends have voted.

Bond et al, 2012, “A 61-million-person experiment in social influence and political mobilization”, *Nature*

- ▶ Costly to measure network structure
- ▶ High overlap across online and offline social networks

OPEN  ACCESS Freely available online

PLOS ONE

Inferring Tie Strength from Online Directed Behavior

Jason J. Jones^{1,2*}, Jaime E. Settle², Robert M. Bond², Christopher J. Fariss², Cameron Marlow³,

bioRxiv preprint doi: [https://doi.org/10.1101/12](#)

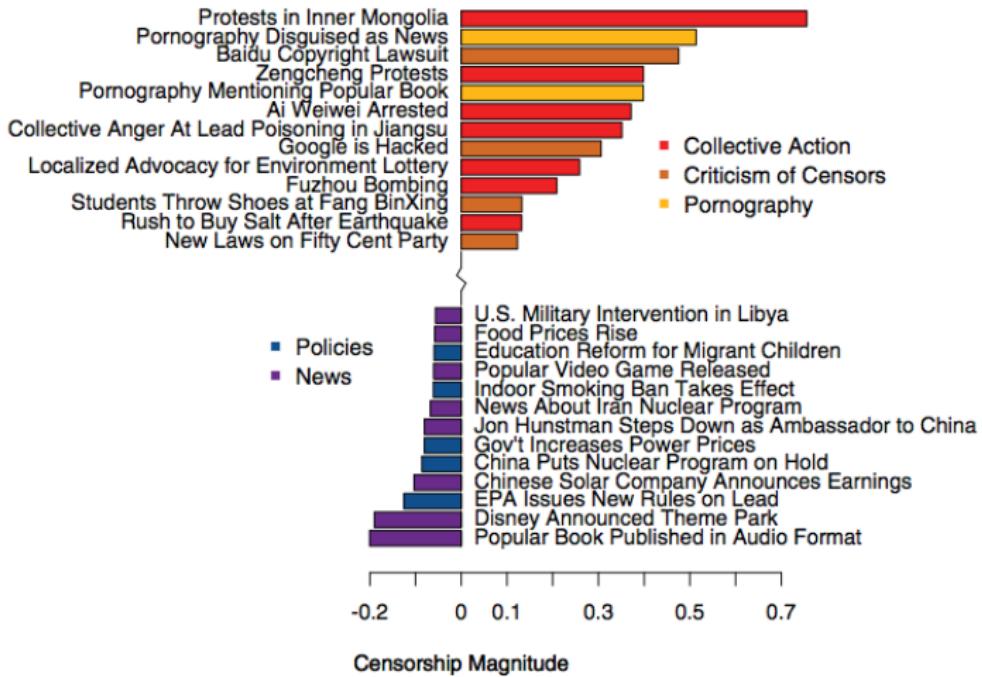
Social media research

Two different approaches in the growing field of social media research:

1. Social media as a new source of data
 - ▶ Behavior, opinions, and latent traits
 - ▶ Interpersonal networks
 - ▶ Elite behavior
 - ▶ Affordable field experiments
2. How social media affects social behavior
 - ▶ Collective action and social movements
 - ▶ Political campaigns
 - ▶ Social capital and interpersonal communication
 - ▶ Political attitudes and behavior

Elite behavior

- Authoritarian governments' response to threat of collective action



King et al, 2013, "How Censorship in China Allows Government Criticism but Silences Collective Expression", *APSR*

- Estimation of conflict intensity in real time

Social media research

Two different approaches in the growing field of social media research:

1. Social media as a new source of data
 - ▶ Behavior, opinions, and latent traits
 - ▶ Interpersonal networks
 - ▶ Elite behavior
 - ▶ Affordable field experiments
2. How social media affects social behavior
 - ▶ Collective action and social movements
 - ▶ Political campaigns
 - ▶ Social capital and interpersonal communication
 - ▶ Political attitudes and behavior

Affordable field experiments



[Political Behavior](#)

... September 2017, Volume 39, Issue 3, pp 629–649 | [Cite as](#)

Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment

Authors

Kevin Munger

Authors and affiliations

Original Paper

First Online: 11 November 2016

2.7k

12k

3

Shares Downloads Citations



13 Sep 2015
@██████████ don't be a n_i g_er



...



Rasheed
@Rasheed██████████

@██████████ Hey man, just remember that there are real people who are hurt when you harass them with that kind of language

Social media research

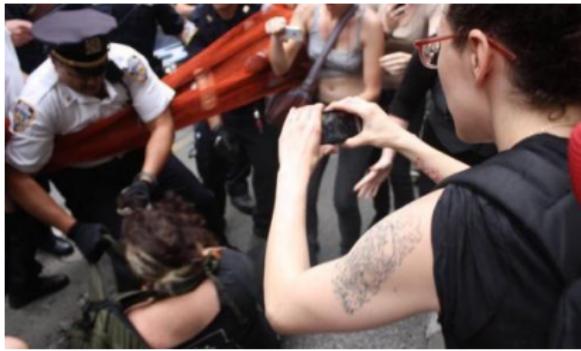
Two different approaches in the growing field of social media research:

1. Social media as a new source of data
 - ▶ Behavior, opinions, and latent traits
 - ▶ Interpersonal networks
 - ▶ Elite behavior
 - ▶ Affordable field experiments
2. How social media affects social behavior
 - ▶ Collective action and social movements
 - ▶ Political campaigns
 - ▶ Social capital and interpersonal communication
 - ▶ Political attitudes and behavior

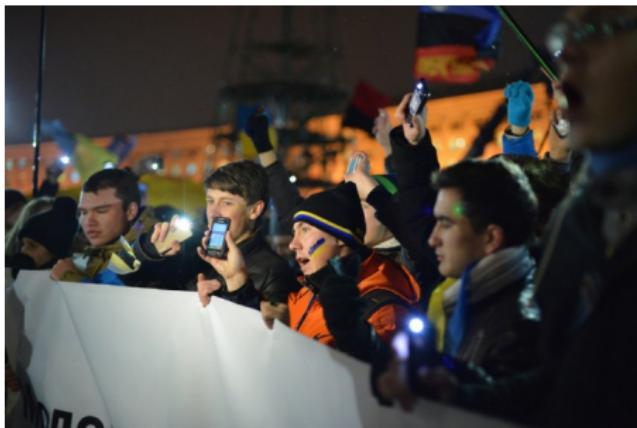




#OccupyGezi



#OccupyWallStreet



#Euromaidan



#Indignados



slacktivism?

Why the revolution will not be tweeted

*When the sit-in movement spread from Greensboro throughout the South, it did not spread indiscriminately. It spread to those cities which had preexisting “movement centers” – a **core of dedicated and trained activists** ready to turn the “fever” into action.*

*The kind of activism associated with social media isn’t like this at all. [...] Social networks are effective at increasing participation – by **lessening the level of motivation** that participation requires.*

Gladwell, Small Change (New Yorker)

*You can’t simply join a revolution any time you want, contribute a comma to a random revolutionary decree, rephrase the guillotine manual, and then slack off for months. **Revolutions prize centralization and require fully committed leaders**, strict discipline, absolute dedication, and strong relationships.*

*When every node on the network can send a message to all other nodes, **confusion is the new default equilibrium**.*

Morozov, The Net Delusion: The Dark Side of Internet Freedom

The critical periphery



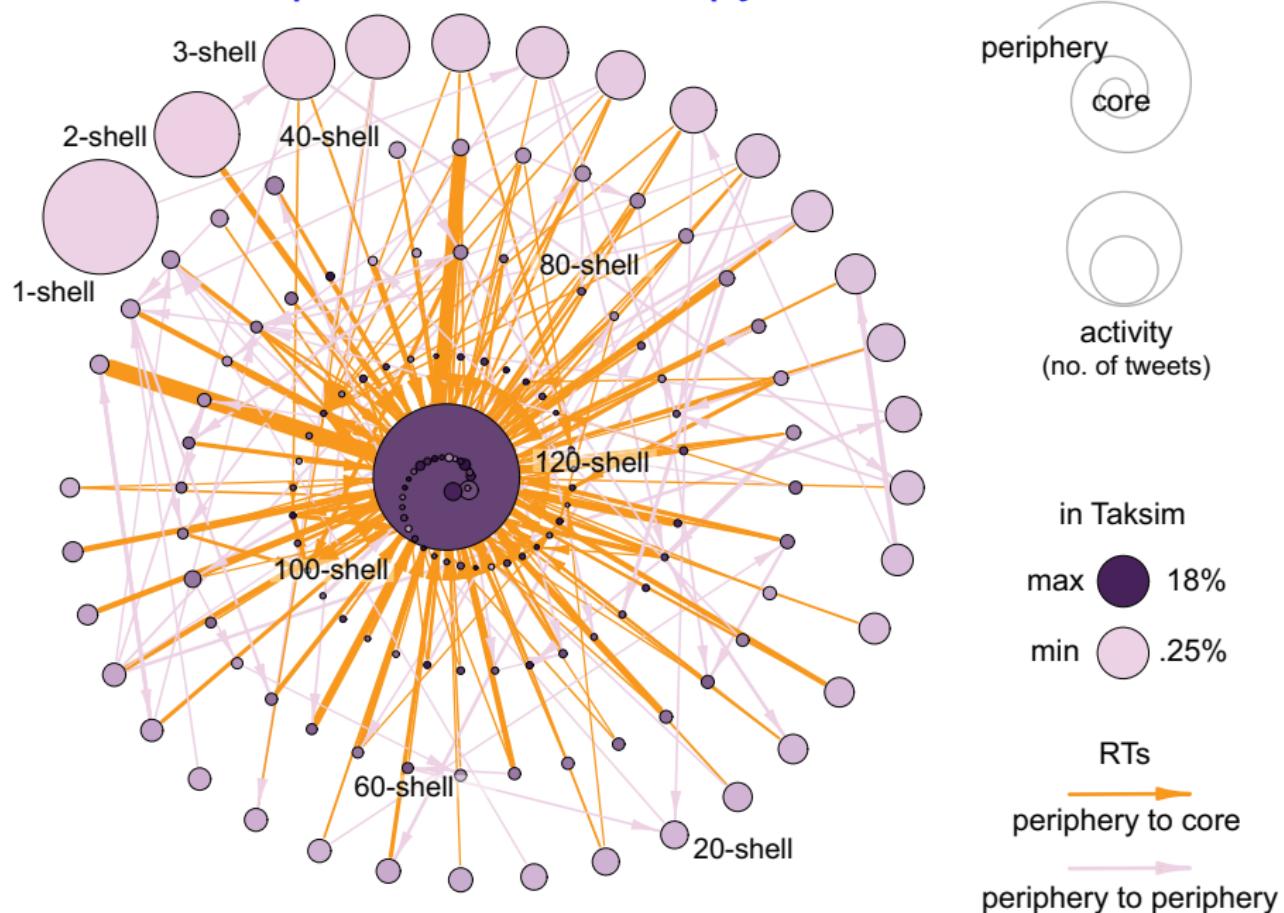
RESEARCH ARTICLE

The Critical Periphery in the Growth of Social Protests

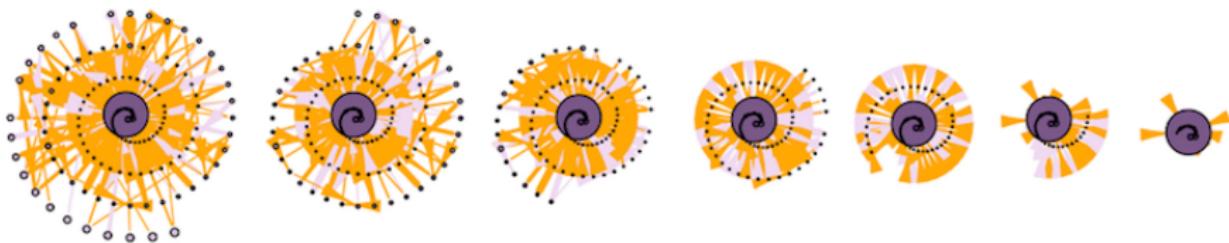
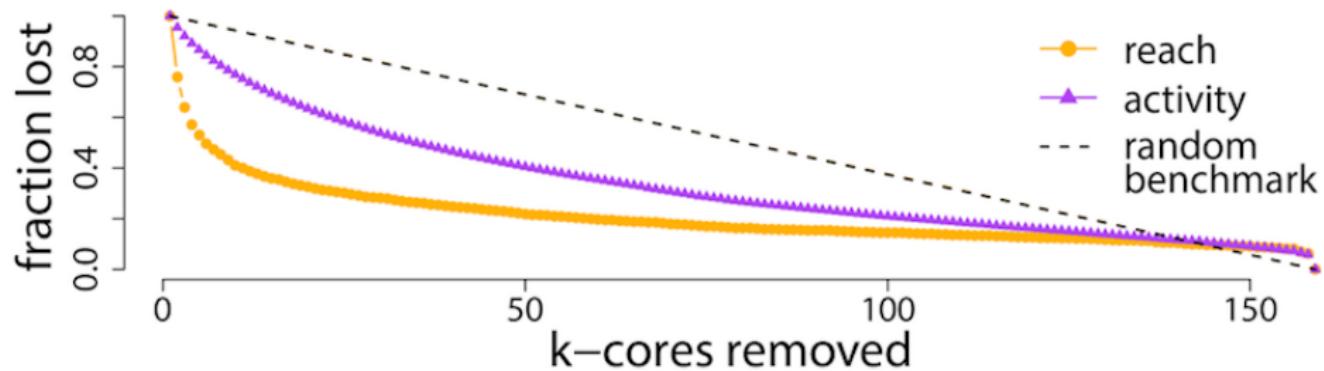
Pablo Barberá^{1*}, Ning Wang², Richard Bonneau^{3,4}, John T. Jost^{1,5,6}, Jonathan Nagler⁶, Joshua Tucker⁶, Sandra González-Bailón^{7*}

- ▶ Structure of online protest networks:
 1. Core: committed minority of resourceful protesters
 2. Periphery: majority of less motivated individuals
- ▶ Our argument: key role of peripheral participants
 1. Increase reach of protest messages (positional effect)
 2. Large contribution to overall activity (size effect)

k-core decomposition of #OccupyGezi network



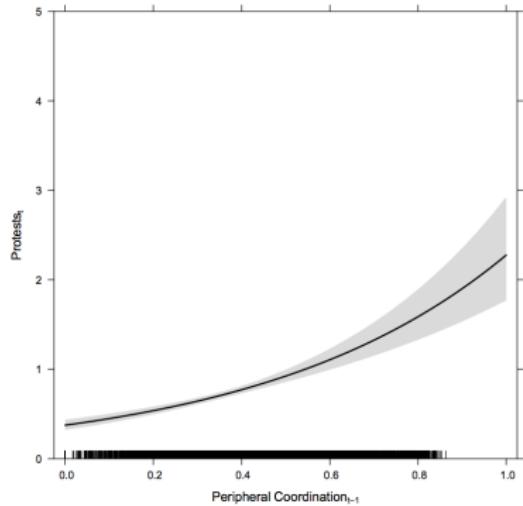
Relative importance of core and periphery



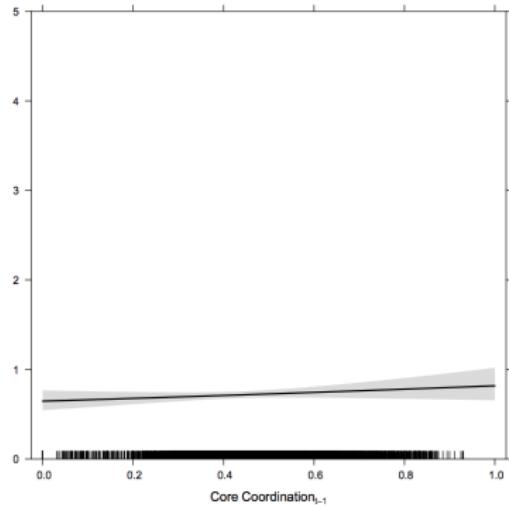
reach: aggregate size of participants' audience

activity: total number of protest messages published (not only RTs)

Peripheral mobilization during the Arab Spring



(a) Increase in protest as peripheral coordination increases



(b) Coordination does not come through core individuals

Steinert-Threlkeld (APSR 2017) "Spontaneous Collective Action"

Social media and democracy

FROM LIBERATION TO TURMOIL: SOCIAL MEDIA AND DEMOCRACY

*Joshua A. Tucker, Yannis Theocharis, Margaret E. Roberts,
and Pablo Barberá*

"How can one technology – social media – simultaneously give rise to hopes for liberation in authoritarian regimes, be used for repression by these same regimes, and be harnessed by antisystem actors in democracy? We present a simple framework for reconciling these contradictory developments based on two propositions: 1) that social media give voice to those previously excluded from political discussion by traditional media, and 2) that although social media democratize access to information, the platforms themselves are neither inherently democratic nor nondemocratic, but represent a tool political actors can use for a variety of goals, including, paradoxically, illiberal goals."

Journal of Democracy, 2017

Social media research

Two different approaches in the growing field of social media research:

1. Social media as a new source of data
 - ▶ Behavior, opinions, and latent traits
 - ▶ Interpersonal networks
 - ▶ Elite behavior
 - ▶ Affordable field experiments
2. How social media affects social behavior
 - ▶ Collective action and social movements
 - ▶ Political campaigns
 - ▶ Social capital and interpersonal communication
 - ▶ Political attitudes and behavior



Barack Obama

@BarackObama



Follow

Four more years.



RETWEETS

756,411

FAVORITES

288,867



11:16 PM - 6 Nov 2012

Sections ≡

The Washington Post

Search



Sign In

Post Politics

**By the end of the 2012 campaign,
every Mitt Romney tweet had to be
approved by 22 people**

Political persuasion

Social media as a new campaign tool:

"Let me tell you about Twitter. I think that maybe I wouldn't be here if it wasn't for Twitter. [...] Twitter is a wonderful thing for me, because I get the word out... I might not be here talking to you right now as president if I didn't have an honest way of getting the word out."

Donald Trump, March 16, 2017 (Fox News)

- ▶ Diminished **gatekeeping** role of journalists
 - ▶ Part of a trend towards citizen journalism (Goode, 2009)
- ▶ Information is contextualized within **social layer**
 - ▶ Messing and Westwood (2012): social cues can be as important as partisan cues to explain news consumption through social media
- ▶ **Real-time broadcasting** in reaction to events
 - ▶ e.g. *dual screening* (Vaccari et al., 2015)
- ▶ **Micro-targeting**
 - ▶ Affects how campaigns perceive voters (Hersh, 2015), but unclear if effective in mobilizing or persuading voters

Social media research

Two different approaches in the growing field of social media research:

1. Social media as a new source of data
 - ▶ Behavior, opinions, and latent traits
 - ▶ Interpersonal networks
 - ▶ Elite behavior
 - ▶ Affordable field experiments
2. How social media affects social behavior
 - ▶ Collective action and social movements
 - ▶ Political campaigns
 - ▶ **Social capital and interpersonal communication**
 - ▶ Political attitudes and behavior

Social capital

- ▶ Social connections are essential in democratic societies, but online interactions do not facilitate creation and strengthening of social capital (Putnam, 2001)
- ▶ Online networking sites facilitate and transform how social ties are established

Tweeting Alone? An Analysis of Bridging and Bonding Social Capital in Online Networks

American Politics Research

1–31

© The Author(s) 2014

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/1532673X14557942

apr.sagepub.com



**Javier Sajuria¹, Jennifer vanHeerde-Hudson¹,
David Hudson¹, Niheer Dasandi¹, and Yannis
Theocharis²**

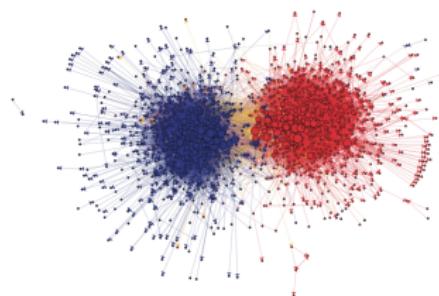
Social media research

Two different approaches in the growing field of social media research:

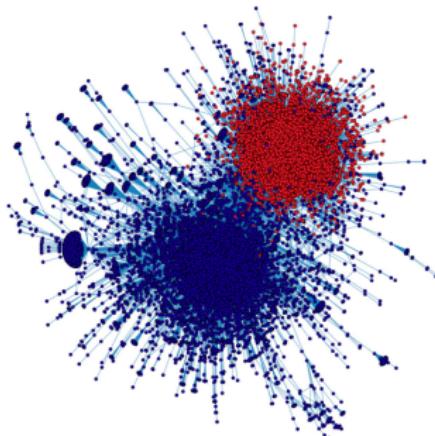
1. Social media as a new source of data
 - ▶ Behavior, opinions, and latent traits
 - ▶ Interpersonal networks
 - ▶ Elite behavior
 - ▶ Affordable field experiments
2. How social media affects social behavior
 - ▶ Collective action and social movements
 - ▶ Political campaigns
 - ▶ Social capital and interpersonal communication
 - ▶ **Political attitudes and behavior**

Social media as echo chambers?

- ▶ communities of like-minded individuals (homophily, influence)



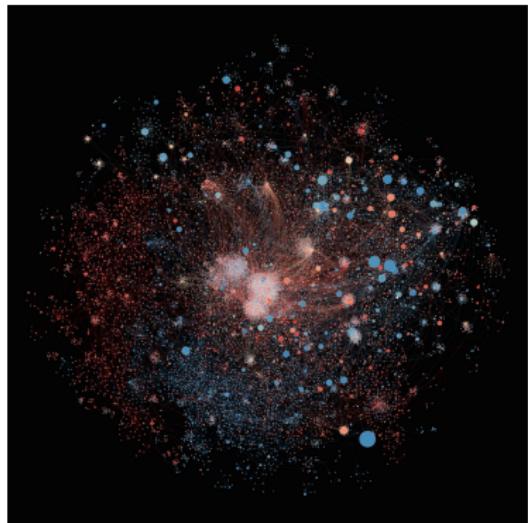
Adamic and Glance (2005)



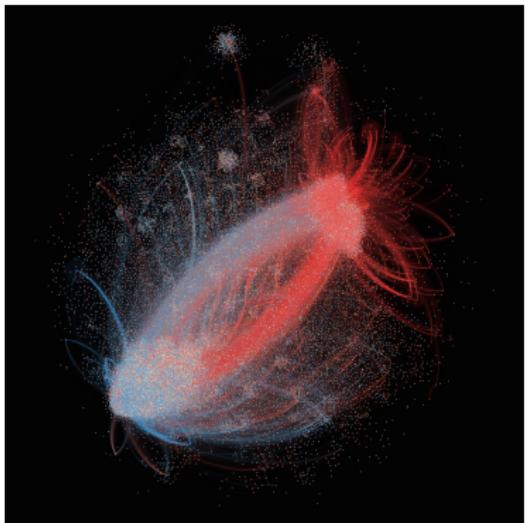
Conover et al (2012)

- ▶ ...generates selective exposure to congenial information
- ▶ ...reinforced by ranking algorithms – “filter bubble” (Parisier)
- ▶ ...increases political polarization (Sunstein, Prior)

Social media as echo chambers?



2013 SuperBowl

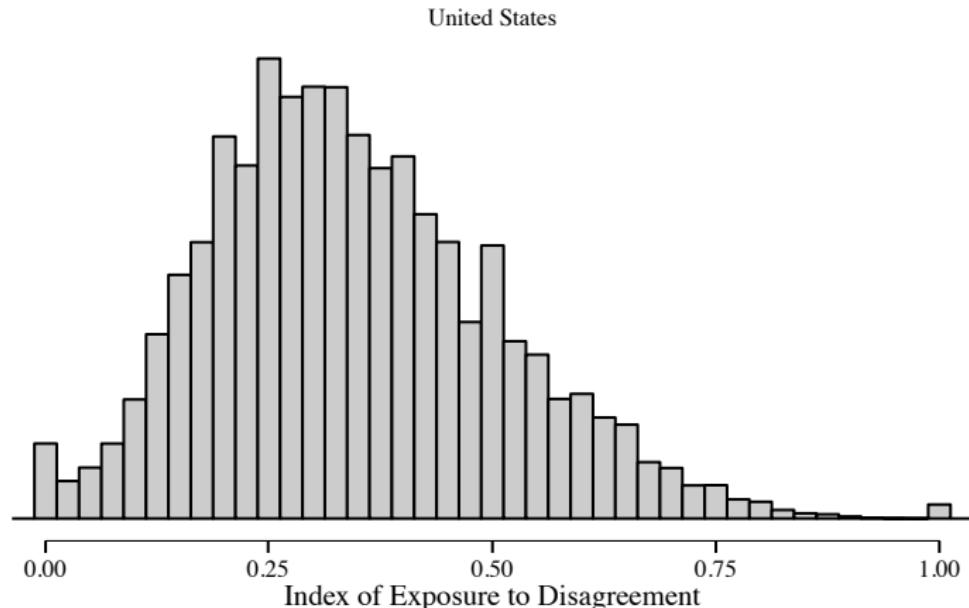


2012 Election

Barberá et al (2015) "Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber?" *Psychological Science*

Measuring exposure to cross-cutting content

Most Twitter users are exposed to high levels of political disagreement

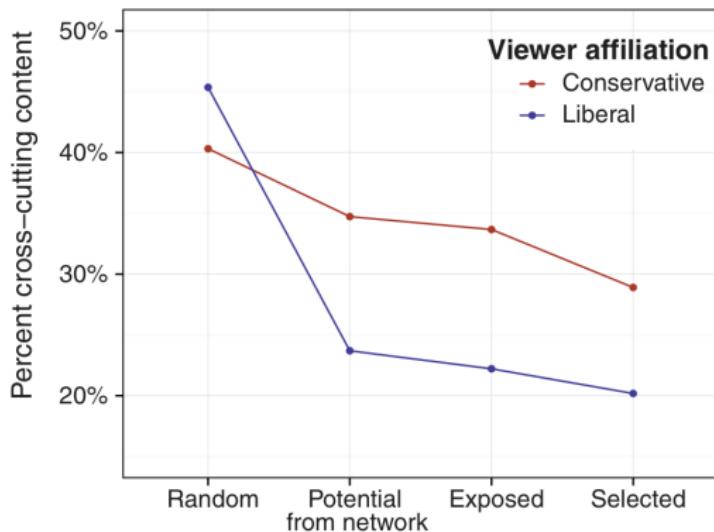


Data: friend networks of ~ 100,000 Twitter users in the US
matched with voter file and following 3+ political accounts

Social media as echo chambers?

Fig. 3. Cross-cutting content at each stage in the diffusion process. (A) Illustration of how algorithmic ranking and individual choice affect the proportion of ideologically cross-cutting content that individuals encounter. Gray circles illustrate the content present at each stage in the media exposure process. Red circles indicate conservatives, and blue circles indicate liberals. (B) Average ideological diversity of content (i) shared by random others (random), (ii) shared by friends (potential from network), (iii) actually appeared in users' News Feeds (exposed), and (iv) users clicked on (selected).

B



Bakshy, Messing, & Adamic (2015) "Exposure to ideologically diverse news and opinion on Facebook". *Science*.

Fake news?



- ▶ Guess et al (2018): **who consumes misinformation?**
 - ▶ Web tracking data: 25% Americans visited fake news websites during the 2016 campaigns
 - ▶ Older, conservative people more likely to be exposed
 - ▶ Facebook key vector of exposure
 - ▶ Fact-check does not reach consumers of misinformation
- ▶ Allcott and Gentzkow (2017): **does it matter?**
 - ▶ Survey experiment with real and placebo fake news stories
 - ▶ Most people do not remember seeing fake news stories
 - ▶ Unlikely to affect citizens' behavior

Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature

By Joshua A. Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan



SHARE



Social media research

Two different approaches in the growing field of social media research:

1. Social media as a new source of data
 - ▶ Behavior, opinions, and latent traits
 - ▶ Interpersonal networks
 - ▶ Elite behavior
 - ▶ Affordable field experiments
2. How social media affects social behavior
 - ▶ Collective action and social movements
 - ▶ Political campaigns
 - ▶ Social capital and interpersonal communication
 - ▶ Political attitudes and behavior

What are the most important challenges when working with social media data?

Social media data and social science: challenges

1. Big data, big bias?
2. The end of theory?
3. Spam and bots
4. The privacy paradox
5. Generalizing from online to offline behavior
6. Ethical concerns

1. Big data, big bias?

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 Pfeffer²

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 inadver-

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 25 to 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 data. Platform-specific sampling problems, for example, the highest-volume source of pub-

The rise of “embedded researchers who have special relationships with providers that give them access to platform-specific data, algorithms, and resources” is creating a diverse media research community. Such researchers, for example, can see a platform’s workings and make accommodations that may not be able to reveal their commercial or the data used to generate their findings.

Ruths and Pfeffer, 2015, “Social media for large studies of behavior”, *Science*

Big data, big bias?

Sources of bias (Ruths and Pfeffer, 2015; Lazer et al, 2017)

- ▶ Population bias
 - ▶ Sociodemographic characteristics are correlated with presence on social media
- ▶ Self-selection within samples
 - ▶ Partisans more likely to post about politics (Barberá & Rivero, 2014)
- ▶ Proprietary algorithms for public data
 - ▶ Twitter API does not always return 100% of publicly available tweets (Morstatter et al, 2014)
- ▶ Human behavior and online platform design
 - ▶ e.g. *Google Flu* (Lazer et al, 2014)

1. Big data, big bias?

Reducing biases and flaws in social media data

DATA COLLECTION

- 1. Quantifies platform-specific biases (platform design, user base, platform-specific behavior, platform storage policies)
- 2. Quantifies biases of available data (access constraints, platform-side filtering)
- 3. Quantifies proxy population biases/mismatches

METHODS

- 4. Applies filters/corrects for nonhuman accounts in data
- 5. Accounts for platform and proxy population biases
 - a. Corrects for platform-specific and proxy population biases
OR
 - b. Tests robustness of findings
- 6. Accounts for platform-specific algorithms
 - a. Shows results for more than one platform
OR
 - b. Shows results for time-separated data sets from the same platform
- 7. For new methods: compares results to existing methods on the same data
- 8. For new social phenomena or methods or classifiers: reports performance on two or more distinct data sets (one of which was not used during classifier development or design)

Issues in evaluating data from social media. Large-scale social media studies of human behavior should i address issues listed and discussed herein (further discussion in supplementary materials).

Ruths and Pfeffer, 2015, “Social media for large studies of behavior”,
Science

2. The end of theory?

Petabytes allow us to say: "Correlation is enough." We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.

Chris Anderson, *Wired*, June 2008

Correlations are a way of catching a scientist's attention, but the models and mechanisms that explain them are how we make the predictions that not only advance science, but generate practical applications.

John Timmer, *Ars Technica*, June 2008

(Big) social media data as a complement - not a substitute - for theoretical work and careful causal inference.

3. Spam and bots



"Follow your coordinators. We need to start tweeting, all at the same time, using the hashtag #ItsTimeForMexico... and don't forget to retweet tweets from the candidate's account..."

Unidentified PRI campaign manager
minutes before the May 8, 2012 Mexican Presidential debate

3. Spam and bots



Ferrara et al, 2016, *Communications of the ACM*

4. The privacy paradox

Online data present a paradox in the protection of privacy: Data are at once too revealing in terms of privacy protection, yet also not revealing enough in terms of providing the demographic background information needed by social scientists.

Golder & Macy, *Digital footprints, 2014*

5. Generalizing from online to offline behavior

What makes online behavior different:

- ▶ Platform affordances may distort behavior (e.g. anonymity encourages vitriol)
- ▶ Tools extend innate capacities (e.g. Dunbar's number)
- ▶ Asymmetries in data availability

6. Ethical concerns

1. Shifting notion of *informed consent*



Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer^{a,1}, Jamie E. Guillory^{b,2}, and Jeffrey T. Hancock^{b,c}

^aCore Data Science Team, Facebook, Inc., Menlo Park, CA 94025; and Departments of ^bCommunication and ^cInformation Science, Cornell University, Ithaca, NY 14853

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 25, 2014 (received for review October 23, 2013)

Emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. Emotional contagion is well established in laboratory experiments, with people transferring positive and negative emotions to others. Data from a large real-world social network, collected over a 20-y period suggests that longer-lasting moods (e.g., depression, happiness) can be transferred through networks [Fowler JH, Christakis NA (2008) *BMJ* 337:a2338], although the results are controversial. In an experiment with people who use Facebook, we test whether emotional contagion occurs

demonstrated that (i) emotional contagion occurs via text-based computer-mediated communication (7); (ii) contagion of psychological and physiological qualities has been suggested based on correlational data for social networks generally (7, 8); and (iii) people's emotional expressions on Facebook predict friends' emotional expressions, even days later (7) (although some shared experiences may in fact last several days). To date, however, there is no experimental evidence that emotions or moods are contagious in the absence of direct interaction between experiencer and target.

On Facebook, people frequently express emotions, which are

2. Most personal data can be de-anonymized

[Ethics and Information Technology](#)

December 2010, Volume 12, [Issue 4](#), pp 313–325

“But the data is already public”: on the ethics of research in Facebook



Shift in communication patterns



Analyzing Social Media Text: First Steps

Why we're using R

- ▶ Becoming *lingua franca* of statistical analysis in academia
- ▶ What employers in private sector demand
- ▶ It's free and open-source
- ▶ Flexible and extensible through *packages* (over 10,000 and counting!)
- ▶ Powerful tool to conduct automated text analysis, social network analysis, and data visualization, with packages such as quanteda, igraph or ggplot2.
- ▶ Command-line interface and scripts favors reproducibility.
- ▶ Excellent documentation and online help resources.

R is also a full programming language; once you understand how to use it, you can learn other languages too.

RStudio Server

RStudio

File Edit Code View Project Workspace Plots Tools Help

Go to file/function

Project: (None)

diamondPricing.R* | formatPlot.R* | diamonds*

Source On Save

1 library(ggplot2)
2 source("plots/formatPlot.R")
3
4 view(diamonds)
5 summary(diamonds)
6
7 summary(diamonds\$price)
8 aveSize <- round(mean(diamonds\$carat), 4)
9 clarity <- levels(diamonds\$clarity)
10
11 p <- qplot(carat, price,
12 data=diamonds, color=clarity,
13 xlab="Carat", ylab="Price",
14 main="Diamond Pricing")
15

15:1 (Top Level) R Script

Console

	x	y	z
Min. :	0.000	0.000	0.000
1st Qu.:	4.710	4.720	2.910
Median :	5.700	5.710	3.530
Mean :	5.731	5.735	3.539
3rd Qu.:	6.540	6.540	4.040
Max. :	10.740	58.900	31.800

> summary(diamonds\$price)

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
326	950	2401	3933	5324	18820	

> aveSize <- round(mean(diamonds\$carat), 4)

> clarity <- levels(diamonds\$clarity)

> p <- qplot(carat, price,
+ data=diamonds, color=clarity,
+ xlab="Carat", ylab="Price",
+ main="Diamond Pricing")

> format.plot(p, size=24)

> |

Workspace History

Load Save Import Dataset Clear All

Data diamonds 53940 obs. of 10 variables

Values aveSize 0.7979

clarity character [8]

p ggplot

Functions format.plot(plot, size)

Files Plots Packages Help

Zoom Export Clear All

Diamond Pricing

Clarity

- I1
- SI2
- SI1
- VS2
- VS1
- VVS2
- VVS1
- IF

Quantitative Text Analysis. Applications to Social Media Research

Pablo Barberá

London School of Economics

www.pablobarbera.com

Course website:

pablobarbera.com/text-analysis-vienna