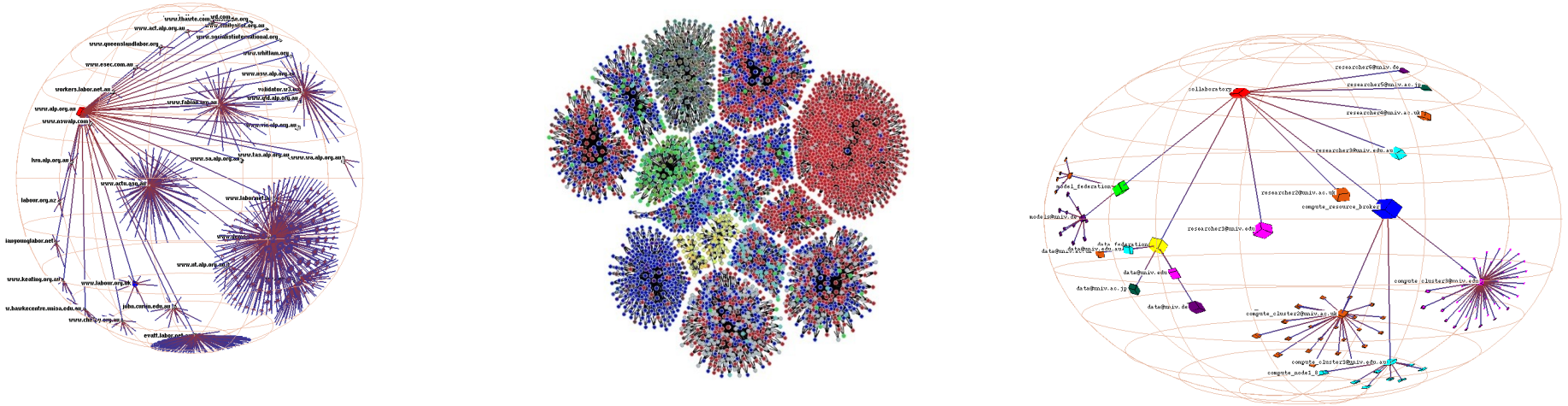


How is Big Data Changing the Nature and Practice of Social Science?



Presentation at ACSPRI Social Science Methodology Conference,
University of Sydney, 19-22 July 2016

Dr Robert Ackland
School of Sociology | Centre for Social Research &
Methods | VOSON Lab
Research School of Social Sciences
Australian National University

E: robert.ackland@anu.edu.au
T: @RobAckland
W: <http://vosonlab.net>



Australian
National
University

ACSPRI | PROGRAMS

2015/16

ACSPRI Methods Courses are intensive small group 'hands on' courses run over five days for researchers and higher degree students.

EACH COURSE COVERS A TOPIC SUCH AS:

- > general statistics
- > multiple regression
- > structural equation modelling
- > experimental design
- > survey research
- > social networks
- > qualitative research
- > mixed methods
- > program evaluation

AND lab based courses use the latest social scientific software and technology. ACSPRI courses cover a variety of levels from Fundamentals (Introductory Level 1) through to very Advanced (Level 5).

2015 WINTER

29th June – 10th July
University of Queensland, St Lucia
20th July – 24th July
University of Western Australia, Perth
Early Bird Deadline 6th May
acspr.org.au/winterprogram2015

2016 SUMMER

18th January – 22nd January
Australian National University, Canberra
1st February – 12th February
University of Melbourne
Early Bird Deadline 18th November
acspr.org.au/summerprogram2016

2015 SPRING

28th September – 2nd October
University of Technology, Sydney
Early Bird Deadline 5th August
acspr.org/springprogram2015

FURTHER INFORMATION

Email: info@acspr.org.au
Phone: **03 8376 6496**
Website: www.acspr.org.au
Find us on Facebook:
[/acspr.org.au](https://www.facebook.com/acspr.org.au)

A full list of courses can be found at www.acspr.org.au/courses and a list for each program is publicised around four months in advance on the Winter, Spring and Summer webpages.

ACSPRI are now offering intensive 2-day weekend Master Classes in our Melbourne office throughout the year. Visit the website for details.

ACSPRI members receive generous discounts on all our courses and master classes.



ACSPRI
Australian Consortium for
Social and Political Research Incorporated

Focus of presentation

- A growing number of social scientists are doing research **about** big data, e.g.
 - Big data cultures
 - Social implications of big data (e.g. surveillance society)
 - Big data governance/policy
- This presentation is aimed at social scientists who are wanting to **do** big data research and who are involved in training next generation of quantitative social scientists
- I deliberately focus on **quantitative research**

Plan

- What is big data?
- Big data and ground truth
- Quantitative social science training (so should I become a data scientist now?)
- The rise of the Application Programming Interface (API)
- A small exercise in big data analysis



What is big data?

What is big data?

- Datasets that cannot be analysed using traditional methods for particular discipline
 - in terms of dataset size, 'big' is relative not absolute (and depends on the discipline) [otherwise the physical scientists would be only people doing big data research...not true]
 - if the data doesn't create imperative to innovate methods, then you probably aren't doing big data research
 - just because dataset is online or digitised (like a lot of our data these days) doesn't mean it is big data

The ABS and big data

- [“Big Data for Informed Decisions: ABS Big Data Strategy,” presentation by Gemma Van Halderen, Population and Education Division, ABS.]
- Big data in context of 'internet of everything' or 'network of networks'
 - **people** (social web) e.g. phone logs/GPS, online social networks (e.g. Facebook), microblogs (e.g. Twitter)
 - **information** (traditional web - the WWW) e.g. web pages, clickstreams, website logs
 - **things** (sensor web) e.g. phones, temperature sensors, medical instruments
 - **places** (geospatial web) e.g. geology, land use maps, weather
- Interestingly, the ABS doesn't refer to census data as big data



Big data and ground truth

What is ground truth?

- http://en.wikipedia.org/wiki/Ground_truth
- In **statistical machine learning** “ground truthing” refers to collecting objective data for training datasets used in supervised learning techniques
- In **meteorology** ground truth refers to data collected on location ...used for calibrating remote sensing techniques (e.g. analysis of satellite imagery)

Ground truth in the social sciences

- Often harder to establish GT in social science, compared with physical/computer sciences
 - behaviour being researched is a construct e.g. existence of social movement, collective identity, strategic action fields
 - behaviour is well defined or can be objectively classified but requires manual coding (domain knowledge) to categorise observations in dataset (e.g. is an organisation pro-choice or pro-life?)

- Nature of GT (how much “truthiness” is there?) is relevant to how big data will impact on social science research
 - Example 1 (GT relatively easy to establish): Using mobile phone records to improve demographich forecasting of migration flows
 - Example 2 (GT not easy to establish): Using Twitter to study “strategic action fields” e.g. Occupy Wall Street:
 - “Constructed mesolevel social order in which individual or collective actors are attuned to and interact with one another on the basis of shared (not consensual) understandings about the purposes of the field, relationships to others in the field (including who has power and why), and the rules governing legitimate action in the field” [Fligstein, N. and McAdam, D. (2011): “Toward a General Theory of Strategic Action Fields,” *Sociological Theory*, 29 (1)]

BIG DATA & SOCIETY



[Home](#) [About the Journal](#) [Bookcasts](#) [Editorial Team](#) [FAQs](#) [Contact Us](#) [Upcoming Big Data related events](#)

Sunday, 9 February 2014

First Volume Contents III: Forthcoming contributions from researchers at Oxford Internet Institute and Microsoft Research

We are pleased to post the following draft abstracts of forthcoming contributions:

RESEARCH ARTICLES

Constructing meaning through big data: Reflexive triangulation and the problem of ground truth in user-generated content

Bernie Hogan, Mark Graham, Ahmed Medhat, David Palfrey, and Ralph Straumann,
Oxford Internet Institute

Big Data & Society



“... **data can never speak for themselves** and [there is] a need for a reflexive consideration of the correspondence between research goals and empirical data. ...when trying to extract meaning from the entirety of Wikipedia’s store of information, we can demonstrate how **there is never any kernel of ground truth** behind layers of user generated content. Focusing on the codified ‘big data’ rather than the processes through which the codification happened runs the risk of concealing the technologies, motivations and cultures that come together to produce this content.”

RESEARCH

Open Access



Disconnected, fragmented, or united? a trans-disciplinary review of network science

César A. Hidalgo

- “I argue that social and natural scientists fail to see eye to eye because they have **diverging academic goals**. Social scientists focus on explaining how **context specific social and economic mechanisms** drive the structure of networks and on how networks shape social and economic outcomes. By contrast, natural scientists focus primarily on modeling network characteristics that are independent of context, since their focus is to **identify universal characteristics of systems** instead of context specific mechanisms.”

Quantitative social science training (so should I become a data scientist now?)

- Most of the interesting data online are socially generated, but given scale/complexity of online data, do you need to be an applied physicist to be able to work with them?
 - *Anecdote: Recent conversation with postdoc at Northeastern University Network Science Institute (applied physicist, now working on computational social science) – I found myself saying: “this is certainly the era for a person with your background”*
 - *Are people saying similar things to postdocs with “traditional” quantitative social science background? Explosion of socially-generated big data should be a boon for social science and creating “jobs for social scientists”*

<https://www.quora.com/How-can-I-become-a-data-scientist-1> (Alex Kamil)

- Learn about matrix factorizations
- Learn about distributed computing
- Learn about statistical analysis
- Learn about optimization
- Learn about machine learning
- Learn about information retrieval
- Learn about signal detection and estimation
- Master algorithms and data structures
- Practice
- Study Engineering

- Traditionally, empirical social scientists did their own data analysis
 - *Anecdote: Indiana University empirical sociologist who successfully collaborates with computer scientists: “previously I would do the data analysis myself but now the scale/complexity is such it is like going into a restaurant and ordering a dish that can only be prepared by a Michelin Star chef”*
 - *What about the social scientists who want to keep cooking for themselves? What are the implications of “contracting out” our data analysis to data scientists?*



The rise of the Application Programming Interfaces (APIs)

- It is fairly easy to work with APIs
 - R statistical software (on CRAN):
 - TwitteR
 - Rfacebook
 - InstaR
 - Similar packages available for python

VOSON SocialMediaLab R Package






- Aims to be the “Swiss army knife” for collecting social media data via free APIs and constructing datasets for network and text analysis
- Current data sources (via free APIs):
 - Twitter (via TwitteR)
 - Facebook (via Rfacebook)
 - YouTube (directly from API)
 - Instagram (via instaR)
 - [...interested in a new data source? We welcome your contribution!]
- Released on CRAN November 2015 – current version is 0.22.0
- More information
 - CRAN page (<https://cran.r-project.org/web/packages/SocialMediaLab/index.html>)
 - VOSON page (<http://vosonlab.net/SocialMediaLab>)
 - GitHub page (<https://github.com/voson-lab/SocialMediaLab>)

Who has contributed to SocialMediaLab?

- **Tim Graham** (Sociology, Univ. of Queensland - soon to be at ANU, @TimothyJGraham) – Lead developer and maintainer
- **Rob Ackland** (ANU, @RobAckland)
- **Chung-hong Chan** (Journalism and Media Studies Centre, Univ. of Hong Kong, @chainsawriot) – implementation of new UI using maggritr



SocialMediaLab data typology as of a month ago...

VOSON SocialMediaLab – Data Typology (5 th April 2016, version 0.20.1)					
	Facebook 	Twitter 	Instagram 	Instagram – Ego 	YouTube 
Data collection	Manually created list of Facebook fan pages	Search on terms (usernames, words, hashtags)	Search on terms in captions OR geographical search (location of person posting caption i.e. uploading photo)	Manually created list of users (who may or may not have posted)	Manually created list of videos
Actor(s)	Users Facebook fan pages posts (not comments)	Users	Users Captions (proxy for photo)	Users	Users
Network(s)	<p>“bi-modal” - directed ties from user to post, based on likes and comments</p> <p>“actor network (users)” - undirected ties from user to user via projection (not yet implemented in package, but via igraph)</p> <p>“actor network (posts)” – undirected ties from posts to posts via projection (not yet implemented)</p>	<p>“bi-modal” - directed ties from user to word/hashtag</p> <p>“actor network” - directed ties from user to user based on @mention, @reply, RT</p>	<p>“bi-modal” - directed ties from user to caption (photo) based on likes and comments. Note that author of caption is stored as vertex attribute.</p>	Directed ties from user to user based on follows	Directed ties from user to user based on mention or reply (“affiliation network”)
Semantic network	No	Yes – words and hashtags are different actor types, edges are co-occurrence in tweet payload	No	No	No
Dynamic network	Yes	No	No	No	No
Text content	Post and comment text Usernames	Tweet payload Usernames	Comment and caption text Usernames	Usernames	Comment text Usernames

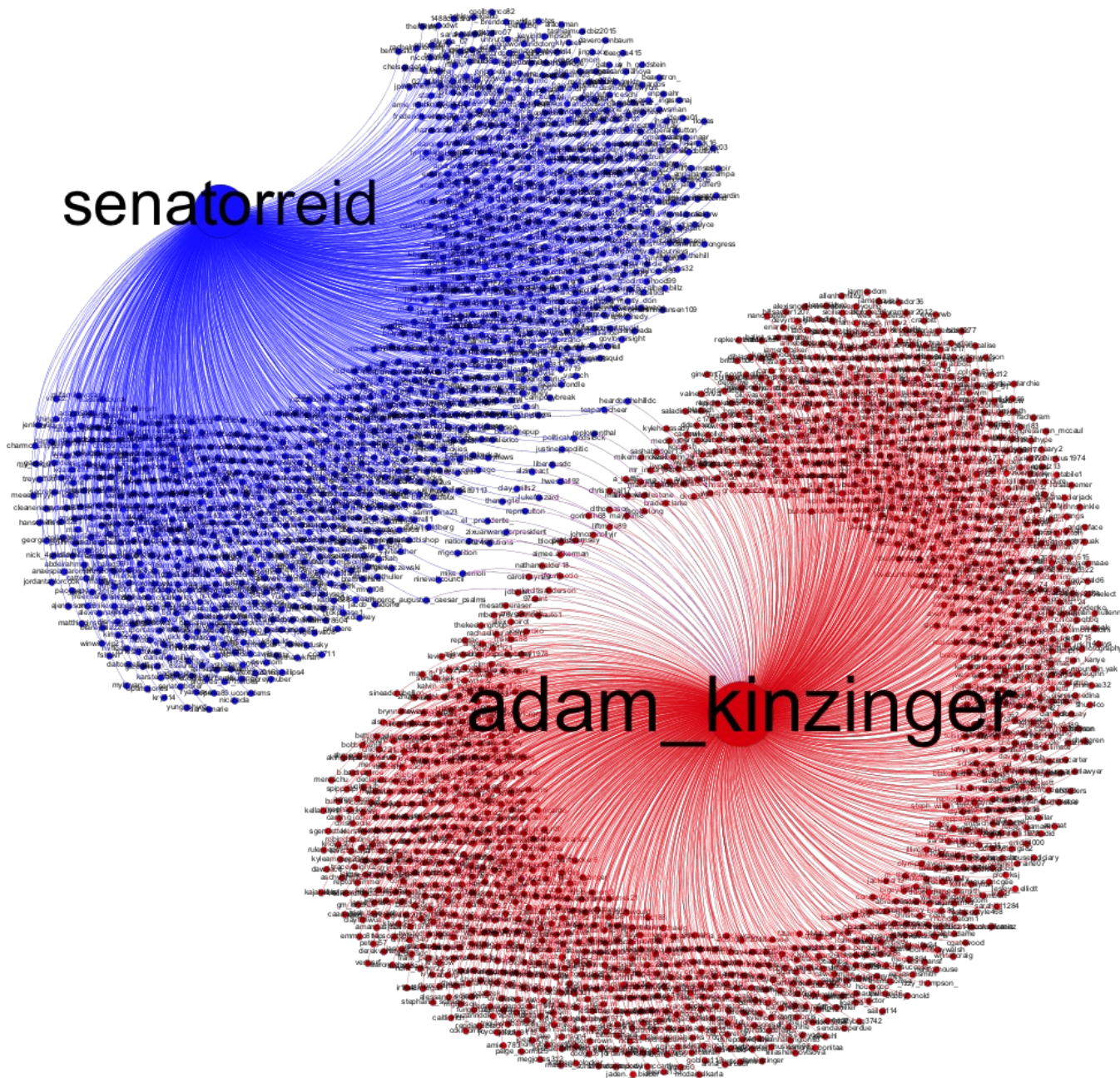
senatorreid

adam_kinzinger

Instagram ego
networks for two
US politicians

Data collected via
SocialMediaLab

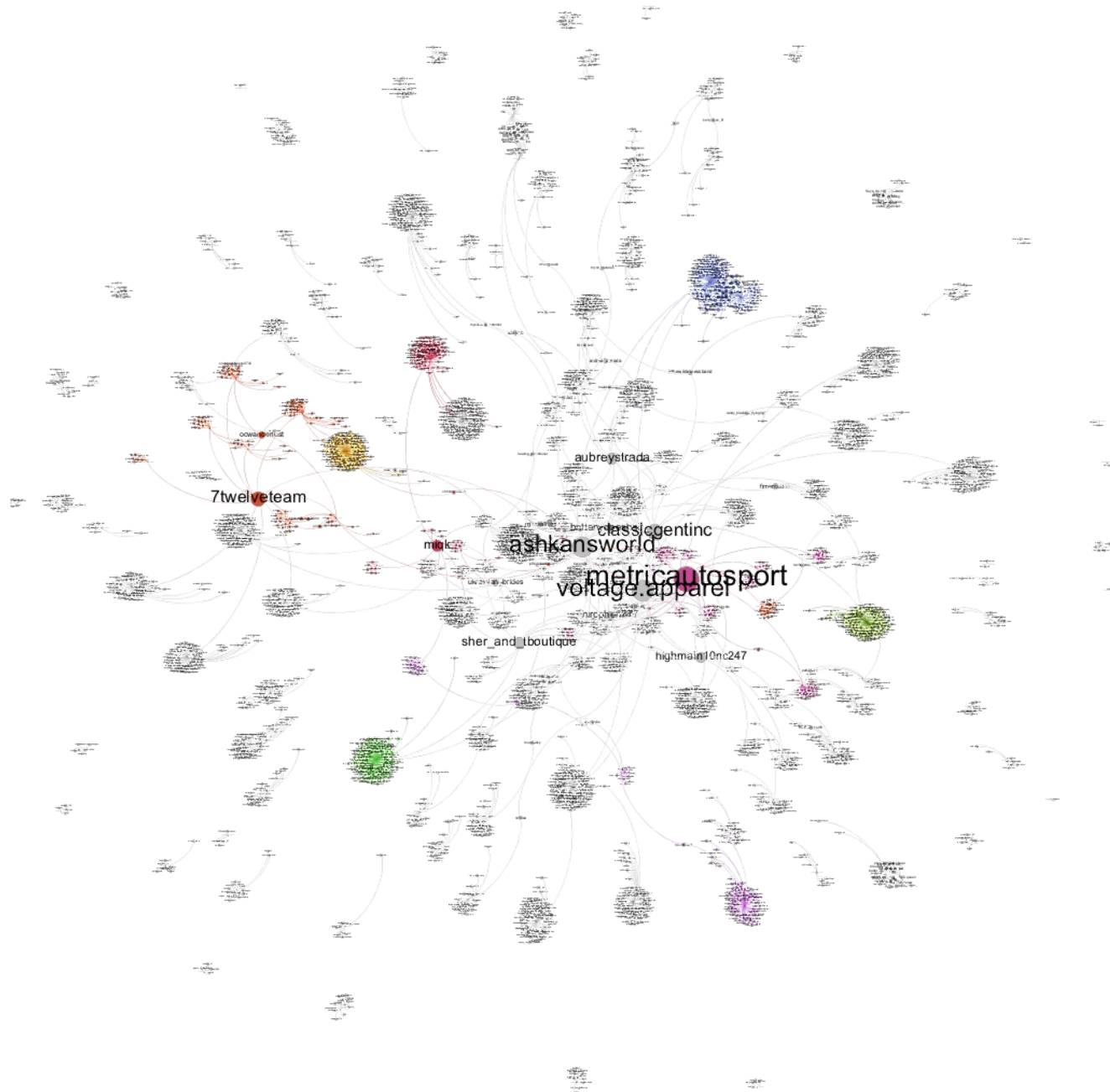
Network visualised
using Gephi



Instagram network in
Newport Beach (site
of the last Sunbelt
International Social
Networks
Conference)

Data collected via SocialMediaLab

Network visualised using Gephi



SocialMediaLab data typology as of today...

VOSON SocialMediaLab – Data Typology (5 th April 2016, version 0.20.1)					
	Facebook 	Twitter 	Instagram 	Instagram – Ego 	YouTube 
Data collection	Manually created list of Facebook fan pages	Search on terms (usernames, words, hashtags)	Search on terms in captions OR geographical search (location of person posting caption i.e. uploading photo)	Manually created list of users (who may or may not have posted)	Manually created list of videos
Actor(s)	Users Facebook fan pages posts (not comments)	Users	Users Captions (prefix for photo)	Users	Users
Network(s)	<p>“bi-modal” - directed ties from user to post, based on likes and comments</p> <p>“actor network (users)” - undirected ties from user to user via projection (not yet implemented in package, but via igraph)</p> <p>“actor network (posts)” - undirected ties from posts to posts via projection (not yet implemented)</p>	<p>“bi-modal” - directed ties from user to word/hashtag</p> <p>“actor network” - directed ties from user to user based on @mention, @reply, RT</p>	<p>“bi-modal” - directed ties from user to caption (photo) based on like and comments. Note that author of caption is stored as vertex attribute.</p>	Directed ties from user to user based on follows	Directed ties from user to user based on mention or reply (“affiliation network”)
Semantic network	No	Yes – words and hashtags are different actor types, edges are co-occurrence in tweet payload	No	No	No
Dynamic network	Yes	No	No	No	No
Text content	Post and comment text Usernames	Tweet payload Usernames	Comment and caption text Usernames	Usernames	Comment text Usernames

- APIs provide benefits (can get data that are hard/impossible to collect otherwise, data arrive in nice formats e.g. json) but there are costs
 - Reliant on third party
 - No one can stop me from scraping public web using VOSON, but Twitter can turn off my access to the Twitter firehose if I violate ToS e.g. publish Tweet payloads (this is why Indiana University's Truthy/OsoMe API does not provide retweet/follows/mentions/replies networks...chill coming from Twitter)
 - API specifications change (e.g. what just happened to Instagram API)
 - Twitter free API: research results can be qualitatively affected by sampling (Gonzalez-Bailon paper)

A small exercise in big data analysis

Three recent “big data” conferences



Social Media & Society

2016 International Conference (July 11-13, 2016 - London, UK)

2016 2nd Annual International Conference on Computational Social Science

IC²S²

24 – 26 June 2016 / General Session
22 – 23 June 2016 / Pre-Sessions

#ICCSS2016

Hosted by the Kellogg School of Management, Northwestern University, Evanston, IL
USA

Complex Networks

FROM THEORY TO
INTERDISCIPLINARY
APPLICATIONS

July 11–13
2016

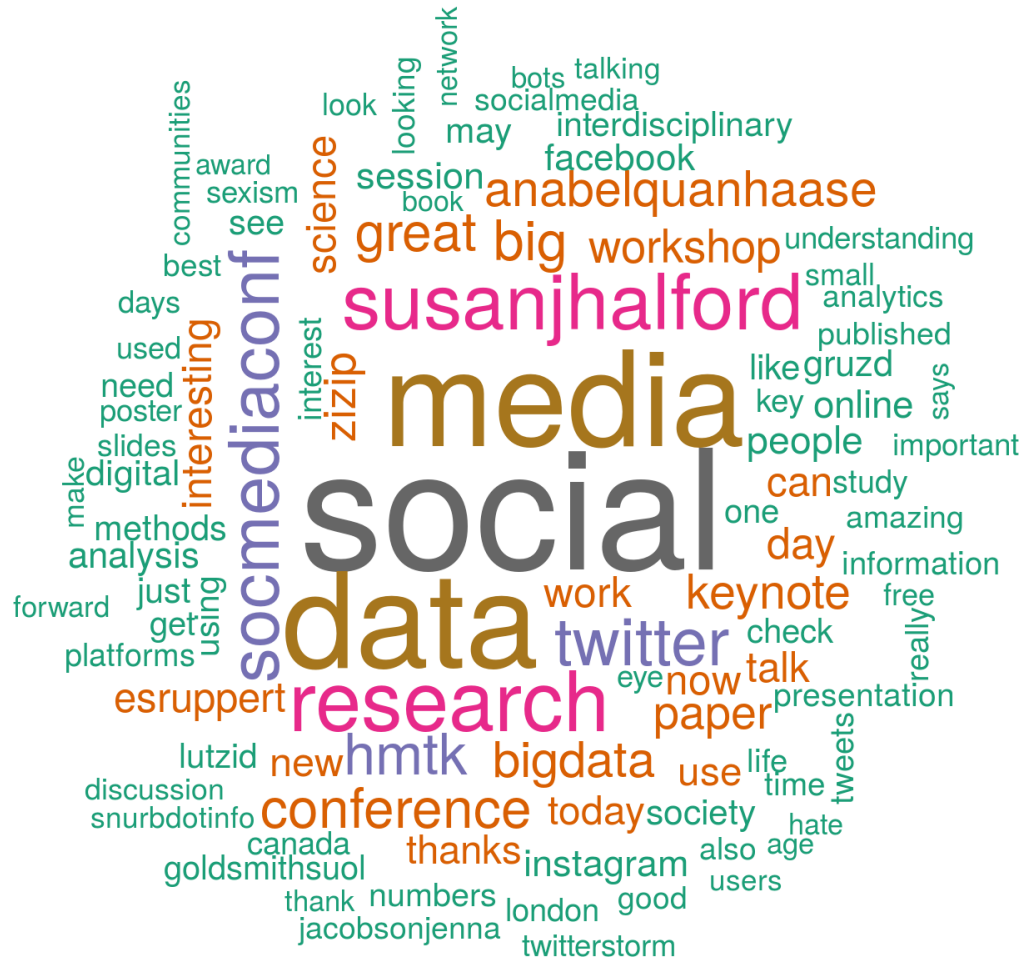
MARSEILLES / FRANCE
Satellite meeting of Statphys26

PROGRAM / COMMITTEES / REGISTRATION / ACCOMMODATION

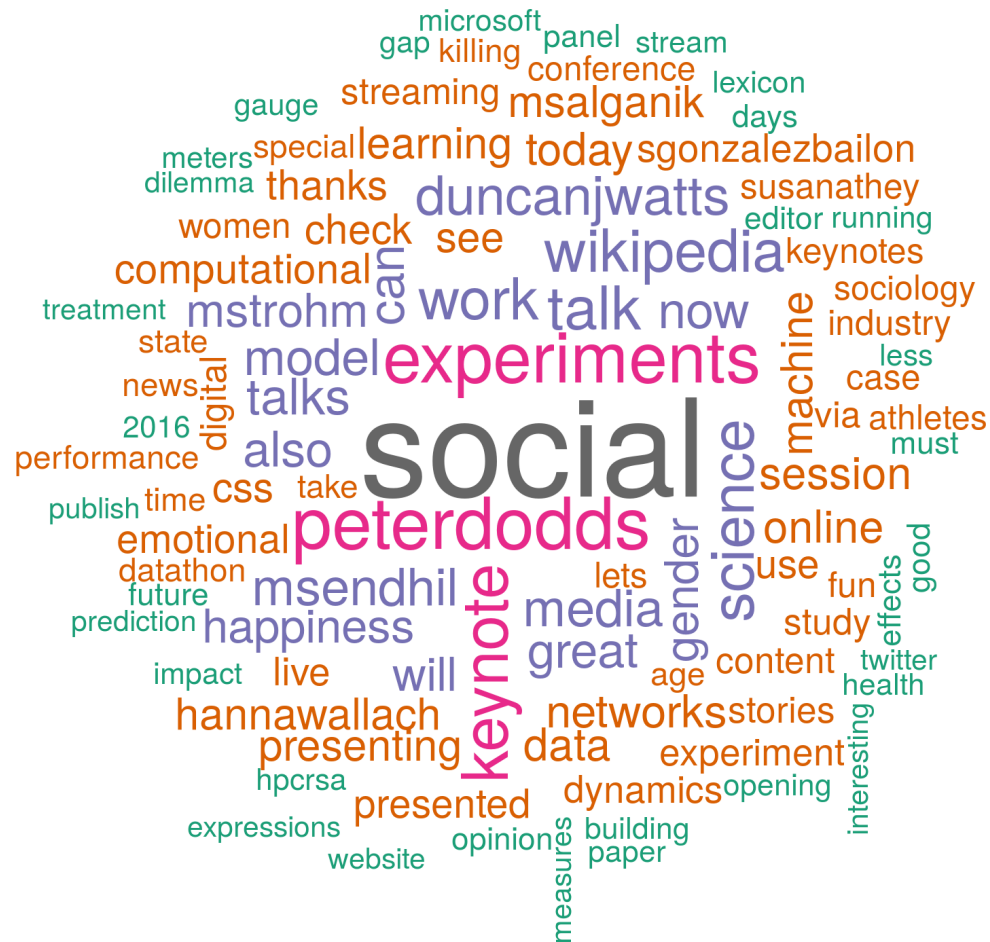
- Can we use Twitter data relating to these three conferences to say something about how big data affecting the practice of social science research?
 - Is there evidence of inter-disciplinarity?
 - Conferences are about investigating the social world using big data: should there be connections/recognition between the tweeters? Are they working in silos?
 - *Anecdote: Recently heard - “when the computer scientists speak at these [Big Data] conferences we [the sociologists] leave the room, and vice-versa”*
 - What topics are they working on?
 - Are social scientists “doing” big data research or only doing research about big data?
- The exercise will hopefully also demonstrate some of the advantages and disadvantages of big data

- Using Twitter Historical PowerTrack API collected all tweets featuring 3 hashtags, authored during period of conference:
 - #smsociety: 4,815 tweets
 - #iccss2016: 572 tweets
 - #complexnets16: 359 tweets
- Identified those twitter users who authored original tweets featuring these hashtags (around 400 users)
- Collected all tweets they authored over the month covering the period of these conferences
 - Approx 450K tweets!
- Analysis:
 - Text analysis of tweet content – tweets authored during the conferences
 - Future work: map retweet/mention/reply network between these 400 users

Word cloud - #smsociety



Word cloud - #iccss2016



Word cloud - #complexnets16



Comparison cloud

smsociety

iccss2016



complexnets16

Commonality cloud



- This was a fairly naive application of big data methods (a bit quick and dirty...), but some reflections:
 - Just because easy (for me) to collect the Twitter data, not necessarily the correct data to answer the question. Perhaps I should use paper abstracts?
 - I chose to use Twitter because I could collect it quickly and easily and get some results, but is the way we should do research?...
 - Are the tweeters representative of the population of researchers?
 - Why did I choose these three conferences/hashtags? Would results be different with other hashtags?
 - I collected these data very easily/quickly, but it would be difficult for someone to replicate this (I can't share the raw data, because of Twitter Terms of Service)



Thank you