Analyzing Big Data with Python PANDAS

This is a series of iPython notebooks for analyzing Big Data – specifically Twitter data – using Python's powerful PANDAS (Python Data Analysis) library. Through these tutorials I'll walk you through how to analyze your raw social media data using a typical social science approach.

The target audience is those who are interested in covering key steps involved in taking a social media dataset and moving it through the stages needed to deliver a valuable research product. I'll show you how to import your data, aggregate tweets by organization and by time, how to analyze hashtags, how to create new variables, how to produce a summary statistics table for publication, how to analyze audience reaction (e.g., # of retweets) and, finally, how to run a logistic regression to test your hypotheses. Collectively, these tutorials cover essential steps needed to move from the data collection to the research product stage.

Prerequisites

I've put these tutorials in a GitHub repository called PANDAS. For these tutorials I am assuming you have already downloaded some data and are now ready to begin examining it. In the first notebook I will show you how to set up your ipython working environment and import the Twitter data we have downloaded. If you are new to Python, you may wish to go through a series of tutorials I have created in order.

If you want to skip the data download and just use the sample data, but don't yet have Python set up on your computer, you may wish to go through the tutorial "Setting up Your Computer to Use My Python Code".

Also note that we are using the iPython notebook interactive

computing framework for running the code in this tutorial. If you're unfamiliar with this see this tutorial "Four Ways to Run your Code".

For a more general set of PANDAS notebook tutorials, I'd recommend this cookbook by Julia Evans. I also have a growing list of "recipes" that contains frequently used PANDAS commands.

As you may know from my other tutorials, I am a big fan of the free Anaconda version of Python 2.7. It contains all of the prerequisites you need and will save you a lot of headaches getting your system set up.

Chapters:

At the GitHub site you'll find the following chapters in the tutorial set:

Chapter 1 – Import Data, Select Cases and Variables, Save DataFrame.ipynb
Chapter 2 – Aggregating and Analyzing Data by Twitter Account.ipynb
Chapter 3 – Analyzing Twitter Data by Time Period.ipynb
Chapter 4 – Analyzing Hashtags.ipynb
Chapter 5 – Generating New Variables.ipynb
Chapter 6 – Producing a Summary Statistics Table for Publication.ipynb
Chapter 7 – Analyzing Audience Reaction on Twitter.ipynb
Chapter 8 – Running, Interpreting, and Outputting Logistic Regression.ipynb

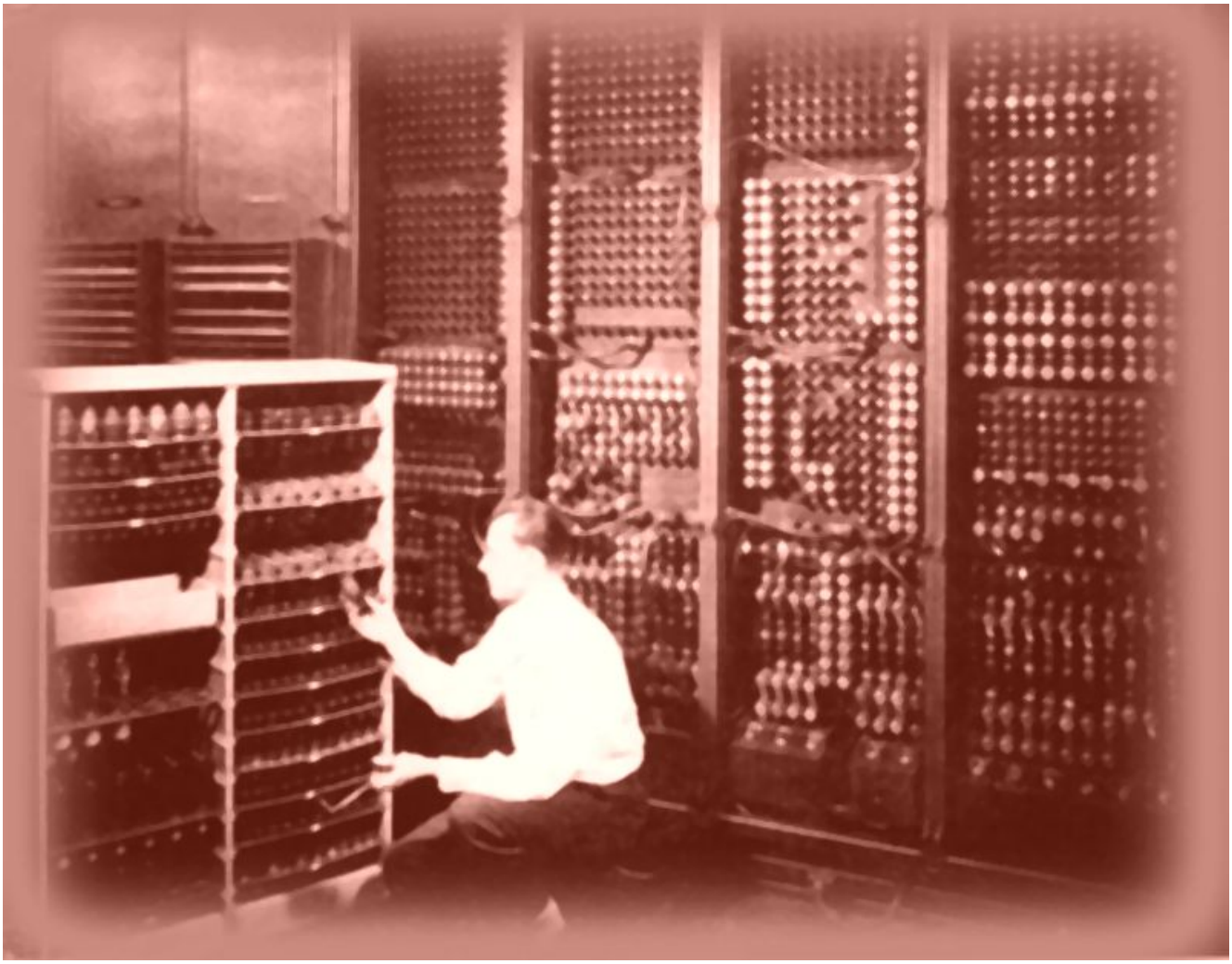
I hope you find these tutorials helpful; please acknowledge the source in your own research papers if you've found them useful:

Saxton, Gregory D. (2015). *Analyzing Big Data with Python.* Buffalo, NY: http://social-metrics.org

Also, please share and spread the word to help build a vibrant community of PANDAS users.

Happy coding!

Do I Need to Learn Programming to Download Big Data?



You want to download and analyze “Big Data” – such as messages or network data from Twitter or Facebook or Instagram. But

you've never done it before, and you're wondering, "Do I need to learn computer programming?" Here are some decision rules, laid out in the form of brief case studies.

One-Shot Download with Limited Analysis

Let's say you have one organization you're interested in studying on Twitter and want to download all of its tweets. You are doing only basic analyses in a spreadsheet like Excel. In this case, if you have a PC, you can likely get away with something like NodeXL – an add-on to Excel. ***VERDICT: COMPUTER PROGRAMMING LIKELY NOT NECESSARY***

One-Shot Download with Analysis in Other Software

Let's start with the same data needs as above: a one-shot download from one (or several) organizations on Twitter. You wish to undertake extensive analyses of the data but can rely on some other software to handle the heavy lifting – maybe a qualitative analysis tool such as ATLAS or statistical software such as SAS, R, or Stata. Each of those tools has its own programming capabilities, so if you're proficient in one of those tools – and your data-gathering needs are relatively straightforward – you *might* be able to get away with not learning programming. ***VERDICT: COMPUTER PROGRAMMING MAY BE UNNECESSARY***

Anything Else

In almost any other situation, I would recommend learning a programming language. Why is this necessary? For one case, let's say you wish to download tweets for a given hashtag over the course of an event. In this case you'll want to use a database – even a simple database like SQLite – to avert duplicates from being downloaded. The programming language, meanwhile, helps you download the tweets and "talk" to the

database. In short, if you are downloading tweets *more than once* for the same sample of organizations, you should probably jump to learning a programming language. Similarly, if you have any need at all for *manipulating the data* you download – merging, annotating, reformulating, adding new variables, collapsing by time or organization, etc. – then a programming language becomes highly desirable. Finally, if you have any interest in or need of medium- to advanced-level *analysis of the data*, then a programming language is similarly highly desirable. ***VERDICT: PICK A PROGRAMMING LANGUAGE AND LEARN IT***

Conclusion

Not *everyone* needs to learn a programming language to accomplish their social media data downloading objectives. If your needs fall into one of the simple cases noted above then you may wish to skip it and focus on other things. On the other hand, if you are going to be doing data downloads again in the future, or if you have anything beyond basic downloading needs, or if you want to tap into sophisticated data manipulation and data analysis capabilities, then you should seriously consider learning to program.

Learning a programming language is a challenge. Of that there is little doubt. Yet the payoff in improved productivity alone can be substantial. Add to that the powerful analytical and data visualization capabilities that open up to the researcher who is skilled in a programming language. Lastly, leaving aside the buzzword "Big Data," programming opens up a world of new data found on websites, social media platforms, and online data repositories. I would thus go so far as to say that any researcher interested in social media is doing themselves a great disservice by not learning some programming. For this very reason, one of my goals on this site is to provide guidance to those who are interested in getting up and running on Python for conducting academic and social media research. If you are a beginner, I'd recommend you work through the

tutorials listed here in order.

How Many Tags is Too Much?

Including a hashtag in a social media message can increase its reach. The question is, what is the ideal number of tags to include?

To answer this question, I examine 60,919 original tweets sent in 2014 by 99 for-profit and nonprofit member organizations of a large US health advocacy coalition.

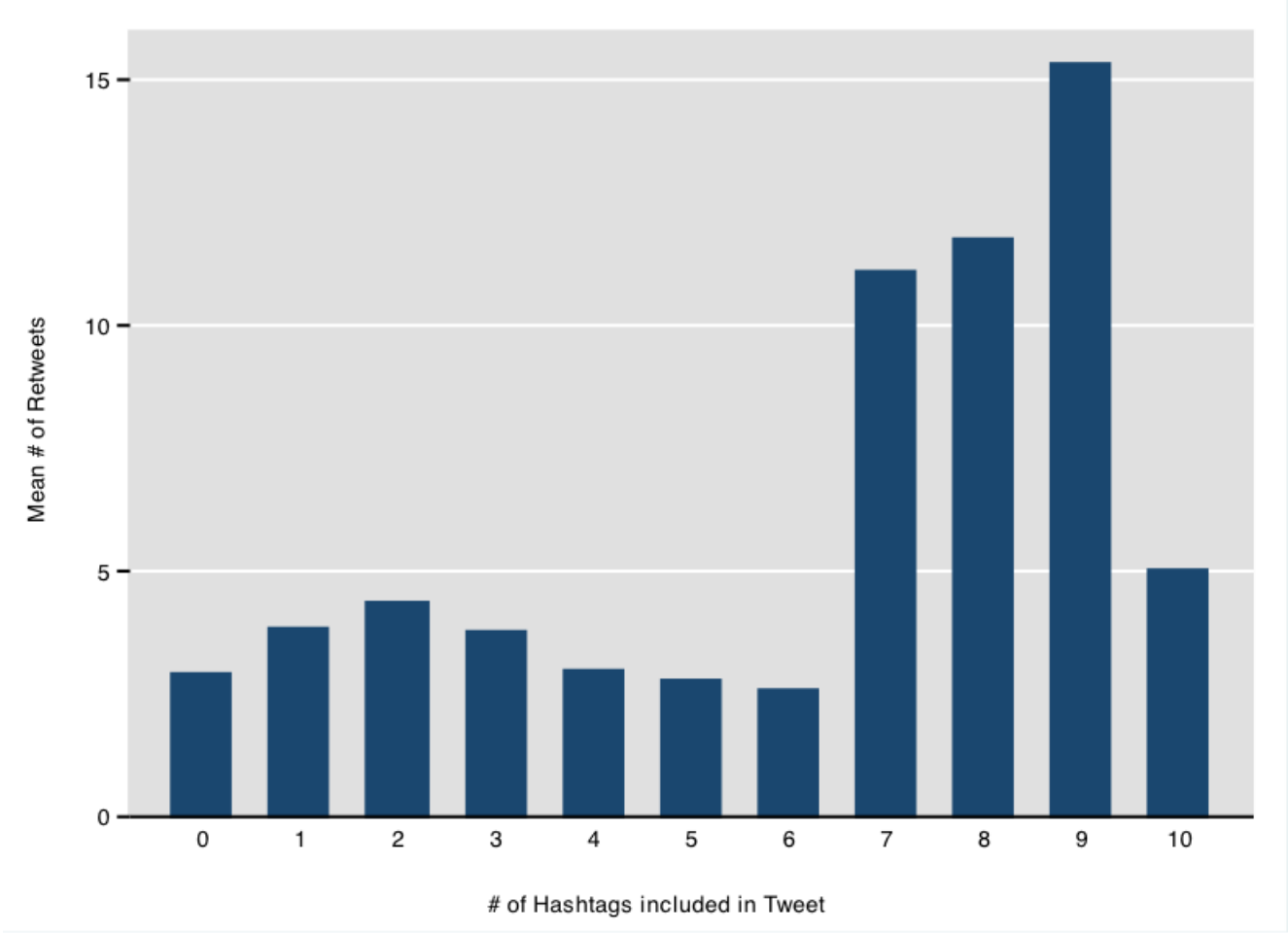
First, the following table shows the distribution of the number of hashtags included in the organizations' tweets. As shown in the table, almost a third (n = 19,747) of tweets do not have a hashtag, almost 39% (n = 23,493) have one hashtag, 19% include two hashtags (n = 11,836), 7% include three (n = 4,381), and 2% (n = 1,161) include 4. Few tweets contain more than 4 tags, though one tweet included a total of 10 different hashtags.

Frequency of Hashtags in 60,919 Original Tweets

# of Hashtags	Frequency
0	19,747
1	23,493
2	11,836
3	4,381
4	1,161

# of Hashtags	Frequency
5	227
6	49
7	13
8	4
9	7
10	1
Total	60,919

Now let’s look at the effectiveness of messages with different numbers of hashtags. A good proxy for message effectiveness is retweetability, or how frequently audience members share the message with their followers. The following graph shows the average number of retweets received by tweets with different numbers of hashtags included.

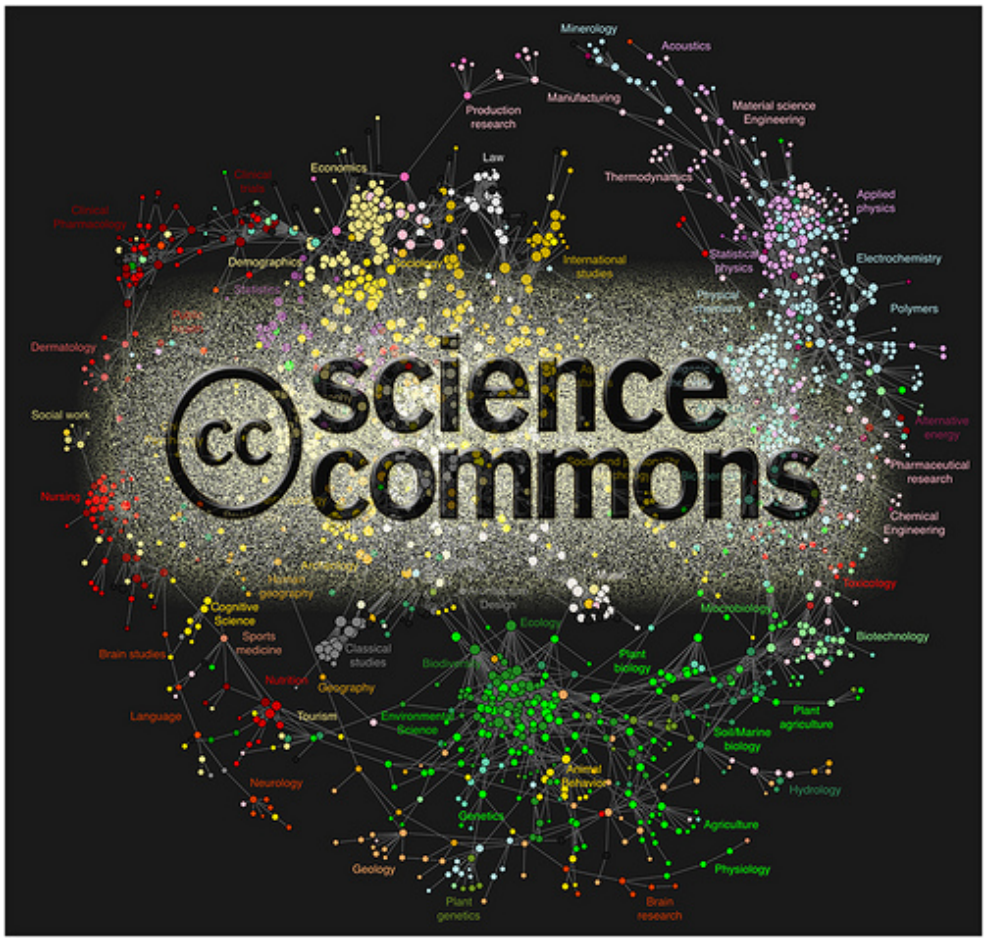


What we see is that more hashtags are generally better, but

there are diminishing returns. Excluding the 25 tweets with more than 6 hashtags, the effectiveness of hashtag use peaks at 2 hashtags, with more than 3 hashtags being only as effective or less effective than no hashtags.

The evidence isn't conclusive – especially given the anomalous findings for the few tweets with 7-10 tags – but there is strong support here that, if you want your message to reach the biggest possible audience, limit your tweets to 1-2 hashtags.

Replication Data



By: Duncan Hull

One of the core tenets of scientific research is *replication.* This gets at the *reproducibility* standard of scientific research. Despite calls for more replication in the social sciences, replication studies are still rather rare. In part, this is the product of journal editors' and reviewers' strong preferences for original research. It is also due to scholars not making their data publicly available. Many of my colleagues in academia, especially those who conduct experimental (lab experiments) research, do not typically make their data publicly available, though even here anonymized data should be available.

By: Seattle Municipal Archives

Replication datasets are not valuable solely for replication studies. In any dataset there are unused variables. A budding scholar or a research-oriented practitioner might be interested in your "leftover" variables or data points. You can't foresee what others will find interesting.

What You Can Do

If you have data, share it. Not only is this being generous, but there is some evidence it may even be good for your career (citations, etc.). If you don't have the capacity to warehouse it yourself, there are archives available for you. A good choice is Gary King's Dataverse Network Project.

My Data

In the spirit of replication and extension, I would like to let people know which data sources I have available. If you'd like any of it, shoot me a message and we'll figure out a way to get it to you.

Spanish Nationalist Event Data, 1977-1996

First, there is a replication archive of data I used in my dissertation. If you're interested in Spanish nationalist contentious politics – specifically, data on violent and non-violent nationalist protests – check out http://contentiouspolitics.gregorysaxton.net/. This site was set up to make publicly available the data used in my dissertation and subsequent publications. There you will find background information on the project, codebooks, data, and copies of articles published using the data. You can browse and search the data and view various interactive graphs. The entire dataset is also available for downloading.

Twitter Data

Twitter data are generally publicly available. However, if you have a pre-defined set of users you can only grab their latest 3,200 tweets, which in some cases is only one year's worth of data. And in other cases, especially if you want to follow a specific hashtag or collect user mentions or retweets, you can only go back one week in time. For this reason, sometimes it can very helpful if someone else has the historical data you

may need. Here are some of the historical data I have, showing the sample of organizations, date range for which data are available, and citations for articles that used the data. If you are interested in it for your own research purposes let me know.

[bibshow file=saxton.bib, format=apa template=av-bibtex-modified]

- *Nonprofit Times 100* organizations – 2009
- *145 advocacy nonprofit organizations* – April 2012
- 38 US community foundations (tweets as well as mentions) – July-August 2011

Facebook Data

Facebook data for organizations is typically public and can be downloaded via the Facebook Graph API. That said, I have some data available on a sample of large nonprofit organizations.

- *Nonprofit Times 100* organizations – December 2009
- *Nonprofit Times 100* organizations – April-May 2013

Website Data

Website data. Historical data can often be gathered from the Internet Archive Wayback Machine, but "robots exclusions" and other errors can prevent this. The following datasets are available:

- 117 US community foundations (transparency and accountability data) – fall 2005 (Saxton, Guo, & Brown, 2007; Saxton & Guo, 2011)[bibcite key=Saxton2007][bibcite key=Saxton2011]
- 400 random US nonprofit organizations – fall 2007 (Saxton, Guo, & Neely, 2014)[bibcite key=Saxton2014]

This is only a partial list of the data I have available. I'll add to this as more data become cleaned and available.

References
[/bibshow]

Establishing a Presence: Advice for PhD Students

Being Director of Graduate Studies gives me plenty of time to reflect on what I'd like students to get out of graduate education. For budding academics, you have all likely heard (countless times!) that the ultimate "deliverable" is high-quality journal articles. Of this there is little doubt – at least in the fields I'm familiar with (social sciences and business). Beyond that, it is important to establish a *presence* in the field. This can involve such traditional activities as reviewing journal articles, presenting and organizing at conferences, conducting guest seminars, and being involved in sub-field specialty groups. With the spread

of new and social media, there is also a new way: *establishing a digital presence*. If you put in the work during your PhD studies, over the course of your graduate career you *will* become one of the world's experts on some area of research, and I would encourage all PhD students to explore the ways that you could make this presence known to your relevant academic community. Increasingly, knowledge and ideas are being shared online – and if you are not actively involved in influencing these knowledge networks you are missing out.

Increasingly, knowledge and ideas are being shared online – and if you are not actively involved in influencing these knowledge networks you are missing out.
I am not talking about just having a LinkedIn or Academia.edu account. Your ultimate goal in establishing a digital presence will be to *add value* to the conversations that are already happening online. This can be done through microblogging on Twitter, Tumblr or LinkedIn, through a conventional long-story blogging platform, or via original video or slide content such as Slideshare. Here are several paradigms for you to consider as you mull over the digital presence that best fits your interests and talents:

- The provocateur – push, prod, and provoke the academy in a direction you feel strongly about.
- The curator – become the source others turn to for by aggregating and re-framing relevant content.
- The teacher – teach others how to do what you know.
- The advice-giver – advice is cheap, but you may have something useful to add.
- The marketer – promote your work, but in a way that is not merely self-serving. Rather, show how your work builds on and enhances existing research. Contribute to the discussion.
- The practice whisperer – translate the findings of your research in a way that practitioners will find useful. Similarly, you could seek to be a *public intellectual*,

as called for by Nicholas Kristof.

These are just some of the ways you can make a presence in the field. Play around with it and find your identity. One of our graduate students, Wayne Xu, has done an excellent job in using a new blog along with Slideshare to take a teaching role. One of UB's long-ago graduates, Han Woo Park, has similarly become one of the most successful posters on Slideshare. Personally, my blog mixes the role of provocateur, teacher, advice-giver, marketer, and translator (I leave the curating to others).

Establishing a digital presence is not a replacement for writing strong journal articles, but it is one of the ways you can make your ultimate impact more powerful. In the end, whether you decide to create a web presence in addition to the traditional route or not, be sure you infuse everything you do with *quality*. The academic world seems huge but it isn't. Word gets around. Lastly, talk with your advisor – he or she is there to help you set a long-term strategy for not only publishing high-quality journal articles but also for making your presence in the field known.

How Organizations Use Social Media: Engaging the Public

The research I've done on organizations' use of social media suggests there are three main types of messages that organizations send on social media: informational, community-building, and "action" (promotional & mobilizational) messages.

Each type constitutes a different way of engaging with the

intended audience:

- *Informational* messages serve to inform – about the organization's activities or anything of interest to the organization's audience. One-way communication from organization to public. The audience is in the role of learner.
- *Community-building* messages serve to build a relationship with the audience through engaging in dialogue or making a network connection. Two-way communication. Audience is in the role of discussant or connector.
- *Promotional & mobilizational* messages serve to ask the audience to do something for the organization – attend an event, make a donation, engage in a protest, volunteer, or serve as an advocate, etc. One-way mobilizational communication. Audience is in the role of actor.

[bibshow file=I-C-A.bib, format=apa template=av-bibtex-modified]

This framework originated in a small "Cybermetrics" graduate seminar I taught several years ago that involved inductive analyses of nonprofit organizations' messages on Twitter (working with one PhD student, Kristen Lovejoy), and Facebook (working with another PhD student, I-hsuan Chiu). This collaborative work resulted in two publications that layed out the basic framework (Lovejoy & Saxton, 2012; Saxton, Guo, Chiu, & Feng, 2011).[bibcite key=Lovejoy2012][bibcite key=Saxton2011b]

Why was this framework innovative or important? Public relations theory had a "relational turn" in the late 1990s, where the focused shifted from an emphasis on strategic one-way communications to building relationships (Broom, Casey, &

Ritchey, 1997; Hon & Grunig, 1999; Kent & Taylor, 1998, 2002; Ledingham, 2003; Ledingham & Bruning, 1998).[bibcite key=Broom1997][bibcite key=Hon1999][bibcite key=Kent1998][bibcite key=Kent2002][bibcite key=Ledingham2003][bibcite key=Ledingham1998] These studies were highly influential and helped re-shape the field of public relations to this date. Around the same time they were published, new media began to take off. The effect was that public relations and communication scholars began to focus on ways organizations were employing relationship-building and dialogic strategies in their new media efforts, contrasting these co-creational and dialogic efforts with one-way "informational" communication. In brief, by the time I started this research there was a substantial body of work on the informational and community-building efforts of organizations on new media.

Yet two key things were missing. One, scholars had yet to examine and code the key tool used by organizations on social media – the actual messages, the tweets and Facebook statuses they organizations were sending. Prior social media studies had looked at static profiles and the like. Two, in focusing on informational vs. dialogic communication, scholars had not recognized the considerable *mobilizational* element of organizations' social media messages. Our study helped build on prior research and fill in both of these gaps. Our inductive study zeroed in on the messages and revealed the substantial use of tweets as a "call to action" for the organizations' constituents, whether this was a call for volunteers, for donations, for social action, for retweeting a message, for attending an event or, indeed, for anything where the organization asked its constituents to "do something" for the organization. We labeled these tweets "promotional and mobilizational" messages or, for short, *action* messages.

I think this "I-C-A" (information-community-action) framework is a useful way of examining organizations' messages, and have

continued to use it in my research on nonprofit organizations, including studies of advocacy organizations (Guo & Saxton, 2014),[bibcite key=Guo2014] of the determinants of social media use (Nah & Saxton, 2013),[bibcite key=Nah2013] and of the effectiveness of organizational messages (Waters & Saxton, 2014).[bibcite key=Saxton2014b]

I am also honored that the framework is also finding itself useful by scholars working in other fields, including those working in the health field (Thackeray, Neiger, Burton, & Thackeray, 2013)[bibcite key=Thackeray2013] and political communication (Xu, Sang, Blasiola, & Park, 2014).[bibcite key=Xu2014]

If you're a social media manager and are wondering about the practical significance of this research, it is important to understand the differences between these different messages, and to have an appropriate mix of each type. Informational, mobilizational, and community-building messages each have a different intended audience orientation that should be tailored to the needs of both the audience and the organization. Don't rely only on the 'megaphone' (informational messages), and don't 'mobilize' (action messages) too often. Most effective will be organizations that actively seek to build relationships with their target audience members. Ultimately, the appropriate mix will depend heavily on the organization's social media strategy – and if you don't have one, you should.

I've created an infographic that shows the differences:

References

[/bibshow]