the WINNOWER

REDDIT

Science AMA Series: We created a map of reddit to make it easier for you to discover new communities. We are Drs. Zachary Neal and Randy Olson from Michigan State University, Ask Us Anything!

ZPNEAL R/SCIENCE

ABSTRACT

Prof. Zachary Neal: I am an Assistant Professor of Psychology and Global Urban Studies at Michigan State University and a Nonresident Senior Fellow at the Brookings Institution. My research uses networks to understand urban and community phenomena at multiple scales ranging from the micro (e.g. neighborhood social networks) to the macro (e.g. global transportation infrastructure), and involves the development of new network analysis methods with a particular focus on bipartite projections (e.g. viewing subreddits as linked by users co-posting behaviors). I also serve as editor of the Journal of Urban Affairs and Routledge's Metropolis and Modern Life book series.

Randy Olson: I run the popular data blog at RandalOlson.com/blog/, most recently known for creating the "Ultimate American Road Trip" and solving *Where's Waldo?*. I tweet daily about data visualization and machine learning at @randal_olson, and moderate the largest online community dedicated to data analysis and visualization on reddit, /r/DataIsBeautiful. Aside from my hobbies, I am an AI and visualization researcher at the University of Pennsylvania (previously Michigan State University) working to usher in the next era of Artificial Intelligence. I do my best to ensure that AI will end up friendly and useful rather than a malevolent Skynet.

We're here to answer all of your questions about our recent work on creating visual, interactive maps of online communities such as reddit to make it easier for you to discover new communities. You can find the reddit map we published here (from mid-2013), along with a writeup on the history and motivation of the project here.

Feel free to hit us with any non-research related questions too. We're here for you to Ask Us Anything!

Update: Thank you all for your questions and comments over the past several hours. We had a blast! We will check back in on this thread over the next few days, but it's time to head back to work. We hope you like our Reddit mapping method, and if you'd like to join the effort to keep the Reddit map updated, we posted a list of our open-source mapping tools and encourage you to get in touch with us.

If you have more you'd like to ask, you can follow up with us on Twitter or email:

Prof. Zachary Neal: @zpneal / zpneal AT msu DOT edu

Dr. Randy Olson: @randal_olson / email

READ REVIEWS

WRITE A REVIEW

CORRESPONDENCE:

DATE RECEIVED:
August 12, 2015

DOI:
10.15200/winn.143886.62183

I'm very interested in data visualization and think this is a very powerful tool, both artistically and for storytelling.

What's the best starting point for someone with little background in coding? Are there any programs or exercises you'd recommend for people to get the feel of data visualization and where to look for the best data?

jobhenrique

I am excited to see that data visualization is growing as rapidly as it is, particularly because it helps tell

ARCHIVED:
August 06, 2015

CITATION:
zpneal , r/Science , Science AMA Series: We created a map of reddit to make it easier for you to discover new communities. We are Drs. Zachary Neal and Randy Olson from Michigan State University, Ask Us Anything!, *The Winnower* 2:e143886.62183 , 2015 , DOI: 10.15200/winn.143886.62183

a better and clearer story with data. But, when it comes to looking for the best data, I am concerned that too often the goal is for the largest or newest or cleanest data, without a particular concern with whether the visualizer really understands the context of the data. The reddit data was very good, and turned out to be useful for creating this navigational map. But, without an understand of what reddit is and how it works, this data – regardless of its other qualities – would have been useless. So, the issue of "where to look for the best data" in part depends on the substantive contexts with which the seeker is most familiar.

Thanks for doing this AMA.

Question: Could you create a Reddit map that would lead one not to like-minded subs, but to subs with different views but same level of seriousness?

The internet in general, and Reddit in particular, make it easy to stay within one's mental comfort zone. In particular, your mapping of Reddit links those subs that share contributors, thus more likely to share similar views. I wonder what metrics you can use to connect subs that differ in views yet discuss similar topics.

By the way, for your reddit interest network, do x- and y- axes have meaning? It would be cool if they did, maybe: x is median length of the comments, y is frequency of new posts.

weaselword
In this map, we link subreddits that share a significantly large number of co-participants. But, there is another way to "link" subreddits and thus create an alternate map for navigation. We could link subreddits that share a significantly *small* number of co-participants, which would yield a map that helps navigate toward subreddits one would be highly unlikely to have discovered independently. I'm less sure about how, in addition to building maps around subreddit similarity/difference, we would also index a subreddit's seriousness.

The map was laid out using the OpenOrd algorithm which, to oversimplify, aims to match nodes' spatial distance in a 2D plane to the number of links separating them in a network. Nodes that are directly connected are placed nearby, while nodes separated by long chains of links are places further apart. In these types of network layouts, the axis can have substantive meanings, much like the axis in a multidimensional scaling plot.. In this case, they likely capture different attributes of subreddits that influence when they are connected or not. But, it is important to note that if there are substantively meaningful dimensions present in our map, they are not necessarily oriented up/down and left/right, and are not necessarily orthogonal.

Hi Drs. Neal and Olson. Thanks for doing this AMA!

Reddit recently banned and 'quarantined' a number of hate-subreddits. The hope is that this will help to prevent hateful ideologies from ruining the reddit experience for the average user (and advertiser).

I wonder if you can comment on how effective you think these policies might be. How interconnected are hate-subreddits? From your experience in modeling communities, does banning effectively suppress repugnant ideas, or do they keep evolving to circumvent bans (a Red Queen effect of sorts)? Beyond other hate-subreddits, do these types of subreddits link to other, less obvious nodes in your model?

SirT6
The meta-communities of subreddits we observe – for example, on sports or on programming – do not actually exist on their own, apart from the subreddits that compose them. Meta-Communities (or, in network science, just "communities) are a phenomenon that emerges from the way that subreddits (or

any kind of node in a network) are linked to one another. Thus, although it may be possible to ban or quarantine individual subreddits, it would be virtually impossible to ban (or indeed even predict) the formation of meta-communities.

Research-related question:

In the article, you mention that smaller subs are often surrounding subreddits that encompass their respective topics on the map. However, it seems like the vast majority of subreddits are often overlapping into other categories, with a small number of clusters far away from anything else.

Does the fact that there are groups of nodes that are completely separated from the rest of the reddit map mean that those communities are good at keeping themselves separated from the rest of reddit? And any subreddit closer to the center means it's more linked (or more likely to be linked) to by users of other, more well-known subreddits?

Such clusters that are further from the center are in the following categories:

- **porn**
- **Certain geographic regions (Iowa, Tennessee, New Brunswick)**
- **My Little Pony**

In the specific case of MLP, the /r/mylittlepony subreddit is quite a bit further to the center than the other subs. In that case, is it essentially acting as a conduit to the other MLP-related subreddits?

Non-research question:

What are some of your favourite cooking recipes?

cheeseburgz
In some cases, subreddits can function as gateways that allow "core" reddit users to navigate their way toward the fringes of the reddit world (or vice versa). The /r/mylittlepony sub might be a nice example of this. In other cases, when a single subreddit or meta-community of related subreddits is far away from anything else, this suggests that the users of the subreddits have particularly confined posting activity. For example, the users who post to /r/Biloxi do not often post elsewhere outside of the meta-community that /r/Biloxi belongs to. As a consequence, it would be difficult for /r/Biloxi posters to navigate their way to other parts of the reddit world (or vice versa) simply by following the lead of other posters they frequently see.

For the casual user of visualizations for work, do you have any recommended reading (besides your own work) for improving the quality of my visualizations, specifically with the limitation of being forced to work in Office?

rugger62
Edward Tufte's work on data and visualization is a classic, and he's certainly had plenty to say about the limitations of Office (in particular, Powerpoint).

How can an ordinary urban planner connect with research like this?

I'm... asking for a friend.

FixinThePlanet
There's a lot of potential overlap between urban planning and network science, but it's been under the radar until fairly recently. I wrote my book "The Connected City" for a general audience to try and

highlight some of the potential. At the micro-level, understanding networks can shed light on how neighborhood communities form or dissolve. At the meso-level, the structure of street networks shapes how different cities (or different parts of the city) are experienced by residents and visitors. And at the macro-level, which has received the most attention, the structure of global transportation and finance networks have important implications for sustainability under the threat of economic or epidemic outbreaks.

What is the aspect or fact that surprised you the most about Reddit while working on this project?

malesurfer
Before starting to work on this project, I really hadn't used Reddit in the past or paid much attention to it. So, over the course of modeling its meta-community structure, I learned a lot of things. But, I was most surprised by the narrowness and specificity of some of the subreddits. That we were still able to detect and visualize a more-or-less coherent meta-community structure despite the number and specificity was, if not surprising, at least reassuring that the world has some order to it.

Hi! Thanks for answering questions. My question is: in what way can cities use data about crime to invent new and novel ways to prevent it or catch the perpetrator?

And in the same way, how could a city use data to foster stronger communities?

Thank you!

dan7899
Cities are already using data to address issues like crime and community-building, though some may be more effective than others. In some of my other work, I've used simulation models to explore the conditions under which strong urban communities form (see: https://www.msu.edu/~zpneal/publications/neal-bigsmall.pdf).

A key challenge is that a data-based solution to an urban problem must also be politically tenable to have a chance at being implemented.

Professor Neal, what classes do you generally teach at Michigan State?

I ask because I am a current sophomore (Psychology and Advertising double major) and I would love to take one of your classes at some point.

SpartansATTACK
Thanks. In Spring 2015 I taught a section of PSY493 on Communities and Social Networks, but most of my teaching is usually graduate seminars on research methods, social networks, simulation models, etc. There are sometimes opportunities for undergrads to register for grad seminars, so keep an eye on the schedule for topics that look interesting, then get in touch with me.

Can you describe how you, or your algorithm, chose to leave out certain subreddits? For instance:

- **several non-quarantined subreddits are not present.**
- **no quarantines subreddits seem to be present.**

so:

1. **Can you clarify what your algorithm omits**
2. **Can you confirm if your are editorializing, and where**

3. Can you state any philosophical bases (e.g. related to editorialization, or otherwise) I tried skimming the paper to look for these answers. Pardon if you state them and I missed it. Thank you.

NewAlexandria
There are two major ways that subreddits can get left out of this map. First, they can be omitted because there was insufficient activity in them. For example, we included only active redditors that had made at least 10 posts, so if a particular subreddit had not been posted in by any "active" redditors, it would not appear here. Second, this map only shows subreddits that shared a significantly large number of co-participants with at least one other subreddit. This likely omits subreddits with fringe, smal, and narrow participant populations.

In our writeup of this project in PeerJ, we avoided engaging in any editorializing. In some places, we speculated on some issues to which our data could not directly speak, but we aimed to be clear about when we were interpreting data/findings vs. speculating.

Go green! What's your favorite flavor of MSU Dairy Store ice cream?

Jeremzz
Go white! I'm not sure how I feel about expanding the Big10 to include Maryland, but I do like the new Terrapin flavor.

Do you have any information on the degree of separation of subreddits?

welhtataum
The 2D map we have been working on is, in part, based on degrees of separation. Subreddits that are linked to one another because they share many of the same posters are located nearby in the map, while subreddits that are separated by many links in the network are located further apart in the map. That is, spatial distance between two subreddits in the map is (partly) a function of the geodesic distance between two subreddits in the network.

Hi Prof Neal & Olson!

I actually have 2 questions. The first is on visualization: It seems that there is always a trade-off between how accessible and how informative data visualization is, especially where complex networks are concerned. It sometimes feels like a lot of 'data analysis' is just turning numbers into colorful diagrams that aren't very useful in disseminating information. What are some new horizons for data visualization that might address these issues, and do you see new platforms of human-computer interaction such as VR as being useful for data representation and manipulation in the future?

My second question is on urban communities and data analysis.. I'm a rising sophomore studying information systems and urban data analysis in a dual major program, but prior to this I was a student of architecture for two years, and I dropped out because I felt like I couldn't solve problems directly enough or on a large enough scale. Realistically speaking, how much practical use does urban data analysis really see, in public policy or civil society community-building initiatives? How can we make a bigger difference, which are the areas that need the most advancement?

freedaemons
How much practical use urban data analysis sees depends a lot on the context. For example urban data analysis is heavily used in decisions about where to locate a new traffic signal, but perhaps is less used in decisions about how to build a neighborhood community. For public policy decisions, so much

of the outcome is shaped by political forces, which unfortunately often have little to do with data or evidence. I think part of the goal of data-driven urban planning is to generate data and analyses that are more useful and accessible, and to work on getting the findings to the right audiences. The best analysis in the world is useless if it's hard to understand, or never reaches the right audience.

Did you purely use posts or comments as well in creating this "map of reddit"? If it's just based on posts, would it not be substantially better being based on comments (or comments and posts)? What assumptions underlie the model for your map?

MurphysLab
Good question. For all 876,961 redditors (circa 2013), we gathered their 1,000 most recent link submissions *and* comments. We counted a particular user as "active" in a subreddit if they had made at least 10 posts. This was necessary both to avoid noise from largely inactive subreddit users, and due to data storage limitations.

How did you become involved with mapping online communities? And is it possible to do this with real life communities?

Aspresso
The field of social network analysis has been mapping offline communities since about the 1930s, when Jacob Moreno first used networks to map relationships among girls at an upstate New York reform school. The primary challenge with mapping offline communities is data collection, which often requires surveying community members.

Hey guys, really cool work you're all doing.

What advice do you have for young people interested in getting into data science fields?

What are some important things to take into account when visualizing data?

PortalGunFun
Thanks. Data science is an exciting field, but I suppose I'd offer two suggestions/pieces of advice. First, it's become so popular that the number of people interested in data science will likely far outstrip the number of data science jobs available. So, consider the viability of a data science career, and explore opportunities to make data science a part of work in a larger field with more (or more varied) opportunities). Second, know your data. To produce really useful data analyses and visualizations requires more than facility with coding and other tools...it requires a familiarity with the substantive context of the data.