SOCIAL NETWORK ANALYSIS IN JOURNALISM:

VISUALIZING POWER RELATIONSHIPS

A Project

presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

---

by

CHEN CHANG

David Herzog, Project Supervisor

MAY 2019

ACKNOWLEDGEMENTS

I want to thank Professor David Herzog, my faculty adviser and committee chair, who has provided me great advice and support since I began to study data journalism. Besides this project, Professor Herzog guided me through graduate school with trust, patience and insight. I also want to thank Professor Mark Horvit and Professor Michael Kearney, who helped me narrow down research angles, polish interview questions and make a concrete research plan. Their advice and encouragement supported me through the whole project.

I also need to give thanks to all the data journalists and designers who shared their valuable experiences and helped me make breakthroughs in my research. Without their help, I would not have been able to include sufficient materials and insights in this project.

I am also very grateful to every instructor I encountered at Missouri Journalism School. Their journalistic ethics, passion and integrity would always shine through my journey in the future.

TABLE OF CONTENTS

FIGURES

CHAPTER ONE: INTRODUCTION

Before I came to the Missouri Journalism School, I studied journalism at Wuhan University. In college, I attended the Dy Club Data Media Lab, participating in online chats given by prestigious data journalists. I also helped translate data journalism projects from ProPublica, FiveThirtyEight, and other news outlets. During several internships at domestic and foreign media outlets, I came to know the importance of data and graphics in the journalism industry.

At the Missouri Journalism School, I worked at the IRE database library and enrolled in various courses to further enhance my understanding about data journalism and strengthen my skills to work as a data reporter. The Computer-Assisted Reporting class taught by Professor David Herzog gave me an opportunity to submit a Sunshine request and negotiate with government departments for public records. I also learned to use SQL (Structured Query Language) to manage and analyze different datasets. In the Investigative Reporting class taught by Professor Mark Horvit, I attempted to combine the data analysis and visualization with narratives and storytelling. In the Advanced Data Journalism class taught by Chase Davis, I learned to use Python to scrape structured data from a website and utilize Python libraries, such as Matplotlib, to visualize large-volume datasets. In the Multimedia Planning and Design class taught by Rob Weir, I learned HTML, CSS and JavaScript to design websites and create interactive graphics. At the IRE database library, I filed Freedom of Information Act requests, used My SQL and R to update databases, checked the integrity of datasets, analyzed databases, and created interactive maps and graphics for clients.

I have developed the knowledge and skill of data visualization and analysis from various courses and jobs, which primes me for conducting this graduate project. During the summer internship at USA Today, I have further learned about the role of data visualization in the journalism industry. My supervisors and coworkers showed me a very good example of maximizing the potential of data visualization and combining journalistic judgment and design aesthetics in a project. I have interviewed various data journalists, developers and scholars regarding the application of social network analysis in journalism. They all generously shared their experiences and lessons, introduced me to tremendous resources and connected me to people with inside knowledge. Based on the literature review and research interviews, I have examined the history, process, limits, challenges and potential of applying social network analysis in the journalism industry. I have also offered some practical advice for data journalists to fully take advantage of social network analysis and keep pace with the cross-disciplinary collaboration in data journalism.

CHAPTER TWO: ACTIVITY LOG

This chapter has included fourteen weekly reports during my internship at USA Today graphics team and ten research reports after the internship.

1. weekly reports

Week 1.

During the first week, I joined the intern orientation and participated in the discussion about what would make a good internship in such a short period of time at USA Today. Our manager Shawn showed us the general structure and workflow in the graphics department the second day after we came to the office.

The main task this week was to conduct research about the mission, strategy, and organizational structure of competitive graphics desks as well as their job titles and descriptions, for example, the Washington Post, the New York Times, AXIOS, Bloomberg, Los Angeles Times, NPR, the Guardian etc.

I have conducted researched about NPR, the Wall Street Journal, the Washington Post, POLITICO and found several challenges as found below.

- Graphics desks sometimes grapple with serving the demands of print sections. To solve that problem, WSJ has embedded graphic designers and developers into other sections in the newsroom. For example, there's one news apps developer who sits near the U.S. editor and develops projects directly with people on that desk. The graphics people who're essentially on that team can develop a deeper knowledge about the subject matter.
- Besides, to foster more creativity and flexibility, WSJ has sent a small group of graphics editors and visual reporters to work outside the news cycle, which is

called the Enterprise Visuals team. Graphics editors have more authority to brainstorm and choose visual projects, without the pressure of the daily news deadlines.

- News nerds are also faced with the bleak prospect of leadership, promotion and salary
- Job-hopping is frequently happening. For example, three NPR visual reporters joined the Washington Post and the Marshall Project in the past year and the NPR visual team was merged into Digital News team, cooperating with video and photo journalists. Several editors from the Guardian US joined the New York Times in just past two years, leaving the US interactive team with only a design editor to lead.

Besides, the staff and organization structure of those competitive graphic/data desks was entered into a spreadsheet for managers to discuss and study, which may help them further adjust the strategy at USA Today.

The research part of my project has not started yet.

Week 2.

During the second week, the print chief Jim Sergent assigned a graphic story about women's suffrage worldwide. Basically, I would be in charge of creating a GIF for social media and an interactive slider to show the historical trend of women's suffrage in the past 100 years.

I have never made a GIF before, so made some research about how to make a GIF efficiently from scratch. Photoshop, After Effects and ImageMagick are popular options in terms of creating GIFs.

I decided to learn ImageMagick based on a step-by-step tutorial from 2017 NICAR Conference, because I attended Lena Gregor's session last year and had a pretty good impression about this tool. A former intern already collected data from different countries and finished creating static SVG pictures in Illustrator. I put together those pictures and GIF through command Line, adjusted the transition time and submitted it to Jim. He offered some advice about adding the missing years to show the passage of time completely and also asked me to add an introductory page at the beginning of GIF as well.

Then, the biggest challenge for me was to create a slider map which allowed readers to from control the time and have a general picture about how women's voting right has been promoted from 1893 to 2017. I decided to use D3 and began to look into its online tutorials, Mike Bostock's blog. I also found similar projects to study its code. Finally, I had at least 4 different versions of design and compared their pros and cons.

The research part of my project has not started yet.

Week 3.

I was dedicated to finishing the slider map this week. One of the biggest challenges was to figure out the color scale in D3. These countries that don't have women's suffrage yet would be filled with grey. These countries that have just granted women's right to vote would be filled with deep blue, but their colors need to be replaced with lighter blue as soon as the slider moves to next year. In this way, as time passes by, more and more countries would become light-blue. This world map will turn from grey to blue in a chronological way. To achieve this goal, I need a higher understanding about the function of color scale and data selection in D3. Then

I dived into the specific chapters in D3’s tutorials and had some related practices several times to figure out the possible route.

After I submitted the draft to Jim, he gave me some effective feedback. Firstly, he offered some good advice about adding some tool tips, so the readers would know which country it is when clicking on different areas. Secondly, the position of the slider was not user-friendly enough. He advised me to put it above the map rather than at the bottom of the page. Therefore, the readers would know immediately the function of this map and start to play around.

Jim has been working at the design desk for newspapers for several decades and was very familiar with user habits. His advice pushed me to try to cultivate a user-oriented thinking pattern with the aim of strengthening the interactivity and boosting the readership. I added some tool tips, changed the position of slider. I also added a “play and stop” button right next to the slider, in order to show the chronological change at a steady speed.

Besides the manual option, readers have another choice to look into the historical change worldwide if they click the play button. The time span was over 100 years, which was kind of long. So I also kept the function of manual control in the slider. The readers could also be able to stop at a certain point and find their interested countries. Then I submitted the draft for further advice. Shawn and Jim didn’t offer further feedback yet.

The research part of my project has not started yet.

Week 4.

This week I was mainly trying to figure out my story ideas and find reliable data sources. I collected all the speeches from Nobel Literature winners in preparation for further text analysis about their nationality, keyword, concerned topic and networked agenda. I was interested in doing a data story about the salary and living status of American writers and the make-up of literature publications in America, and reached out to some non-government organizations for help in seeking datasets, but didn't hear back yet.

Another story I wanted to purse was to analyze the demographic characteristics and growing trend of Chinese political prisoners using the data collected by Senator Marco Rubio's team. I conducted some initial analysis in R to get a full picture about the basic shape and potential angles. I have also been trying to produce my own story regarding the border separation or immigrants' children. I considered collecting some hot topics and communication patterns using the Twitter scraping tool. I tried the rtweet package for interacting with Twitter's REST and Stream APIs in R and followed Professor Mike Kearney's tutorials at NICAR Conference, but I didn't persist in pursing this story at last. I should try to practice the network analysis during my internship, since it is also related with my graduate project.

The research part of my project has not started yet.

Week 5.

This week was kind of slow in the graphic desk as well, since it was Independence Day Holiday. I was caught in a fever and mainly worked from home.

Last week the developer Pim Linders told me to test the slider on mobile screens and some other different platforms, such as Mac and Windows. It is actually a key area I have never really dived into before. But it is indeed crucial, for most of the products in newsrooms have been increasingly mobile-driven in recent years. I looked into some tutorials online about mobile-friendly design, tested my graphic using Virtual Box, focused on debugging the slider , inspecting with Chrome's developer tool and tried to make it more adaptable on screens with different sizes.

I also began to make research about Chinese political prisoners and downloaded some papers about the historical background and policy change. I was mainly interested in the geographical distribution, criminal charges, ethnicity, religion and the general historical trend. I looked into the previous news coverage, which was mainly limited to the breaking news and individual stories and lacking in a more comprehensive data analysis about this group. I found a story by Huffington Post that involved some quantitative analysis, which is a good example for me to learn from.

The research part of my project has not started yet.

Week 6.

I submitted all the code about Women's suffrage map on Trello and waited for further guidance. This week I also proposed a story about the supreme court judges. I planned to make a stream graph to show the historical change of religious and ideological shift in the Supreme Court. Since the 1990s, there has been a striking religious shift toward Catholics and Jews among the supreme court judges, in a contrast with the Evangelicals' increasing influence in Trump's administration and the Protestants' dominant role in U.S. adults(48%)and Congress(55.9%). Although there's not a direct relationship between their judicial decisions and religions/ideologies, their background information may help illustrate the polarized divide and predict the potential change on some fundamental rulings.   I collected some related news articles for background research and browsed lots of scholar's papers and databases:

- religion and other biological data:

Free Law Project.

-Lists of all the judges but the religion data is missing:

Supreme Court official website

The Green Papers

-ideology measurement: MQ and JCS scores are both authoritative,

1.Martin-Quinn scores(1938-2016): based on judicial voting; cited by NYT1, NYT2, Axios

2. Judicial Common Space Scores(1937-2016), based on the ideology of the president who appointed the judge and the senators from the judge's home state; cited by Axios

3 Clerk-based Ideology Scores: based directly on the judge's own political donations. Much less used.

After comparing the pros and cons, I decided to use Martin-Quinn scores from Lee Epstein, Andrew D. Martin and Kevin Quinn research to start my graphic.

The research part of my project has not started yet.

Week 7.

The Supreme Court story improved my skills in data manipulation with R. I downloaded all the datasets about Supreme Court justices from Professor Lee Epstein's website. But the original data frame was very tricky and unsuitable for visualization. I used "spread" and "gather" function of R to change the data frame and used "fuzzy joins" to combine the religion dataset and ideology dataset. These two sets have used different styles to record judges' names. So I cleaned the name format in Excel at first. I learned to use “mutate” in R to add columns and merge datasets, which was a very handy function.

This week, I also got engaged in the group meeting of the long-term, multimedia project “Toxic City” regarding the air and chemical pollution along the Mexico-United States border, which is produced by the graphic desk, USA Today Storytelling Studio and Virtual Reality desk. I made sliders about the brainstorming process and helped highlight some key points to be focused on in the design process. During this process, I learned about the creative process at the first stage of a large-scale, long-term project:

1. Read: Go through all available drafts;

2. Cluster: Organize draft content into general categories;

3. Key points: Tease-out key ideas and topics under each category;

4: Selection**:** Voting exercise to rank ideas and topics;

5. Top 3 ideas: Identify from each category based on votes;

6. Top 1 idea: Identify from each category based on votes;

7. Final exploration: Identify any last possible brainstorming sessions based on findings.

The research part of my project has not started yet.

Week 8.

This week, I focused on learning D3 to create an interactive project: a multi-series line chart about the ideological lean and a streamgraph about the religion makeup of Supreme Court justices. And it would be ideal to add a toggle button to allow readers to switch different graphics and make comparisons. Initially, for the ideology chart, a big challenge for me was about preparing datasets in D3. In order to show how each justice changed their ideology over time, I categorized the dataset by year and justice. In this process, I enhanced my understanding and skills about "D3.csv" to get the data set-up to be passed to D3 to render the SVG paths.

I added a "g" element for each series, which allows us to group other items like a text label if we want later, and then passed its "values" array to "line".

Similar with the multi-series line chart, the streamgraph is made up of paths for each of the series in the data. Instead of drawing lines, the paths create an irregular shape that fills the area taken up by the series. To draw the path, I made use of the d3.layout.stack() and d3.svg.area() tools, which could help figure out the coordinates of each "layer" or series of the streamgraph.

I also had a hard time adjusting the style of legend and axis as well. I studied more about the "function" part in JavaScript in order to trigger certain events by certain actions, such as "click", "mouse over".

In total, I learned about how to make an interactive multi-series line chart, area chart and bar chart in D3.

The research part of my project has not started yet.

Week 9.

I finally finished the interactive draft of the Supreme Court justices and submitted it to my manager Shawn. He was occupied with another group project, so we didn't have chance to talk about the details.

I further inspected similar projects created by other news agencies such as the New York Times and Axios and critiqued my own project. I found that the biggest deficiency would be the lack of simplicity and originality. I added too many extra details in the tooltip, because initially I was afraid that this graphic may not fully contain the key information from my raw dataset.

But the addition of those details made this project kind of distracting and annoying when you first opened the webpage. The readers may find it hard to pore over 10 religions on the right side of the graphic and may spend a long time learning the legend. I asked another intern to help me critique my project as well, in a way the classmates mutually evaluated each other’s web design in the Multimedia Planning and Design class at Mizzou. He offered some advice to choose a darker color scheme, since the Supreme Court story would be corresponding to a more serious or solemn style. I deleted the unnecessary columns in my data set, further simplified my design, optimized the color scheme, and focused on the essential information I wanted to deliver to the readers.

I also made a call with my instructor professor Herzog and we discussed the defense date. The research part of my project has not started yet.

Week 10.

This week I was mainly focusing on learning D3 and R to meet the demands of my project. I found I had a very weak foundation in understanding JavaScript. Most of my techniques come from Google search and W3schools tutorials. Previously, whenever I found a problem I could not solve myself, I usually turned to Stack Overflow and copied the solution or code from other people. On the one hand, it was a quick and efficient way to debug my project and move on to the next one. But on the other hand, my understanding about JavaScript was stagnating at the superficial level.

I listed most of my confusions on paper and tried to pick up the basic concepts from scratch. I checked some textbooks such as Eloquent JavaScript, JavaScript: The Good Parts by Douglas Crockford, You Don't Know JS by Kyle Simpson, and followed the practice or finished the assignments in certain sections, in order to strengthen my understanding about "function call", "loop", "object".

As for the research part, I also started contacting ICIJ for my research and interview. The lead developer Pierre Romera was on a vacation until August 24. I'll contact him again at that time.

Rocco Fazzari, designer of the Influencers project, replied to me that he would be happy to take the interview and would let me know the best time later. I've also sent requests to several other interviewees and waited for their response.

Week 11.

This week I was focusing on the introductory text of the women's suffrage story and supreme court story. I looked into some research about the timeline of women's voting right in the world and the relationship between the religion or ideology background of justices and their voting behaviors. I extracted the essential information and further polished my graphics. I attended the farewell dinner to see another intern off. I also attended the intern meeting with Chief People Officer Dave Harmon and have got some valuable advice about personal development and career plan.

This week I've also made some progress in scheduling the research interviews with a couple of reporters and designers. Sarah Cohen sent me a story she did for the New York Times regarding the unregulated background check over the gun purchase online. Professor Brant Houston scheduled an interview with me next Tuesday. Reg Chua from Reuters would take the interview next Friday. Irene Jay Liu from Reuters hasn't replied to my request yet. William Lyon, developers from ICIJ travelled to Europe next week.

I've also got hold of Ben Fry and Mark Schifferli, designers of Connected China. They were happy to talk about Connected China with me.

I've spent some time reading my proposal and looking into some basic concepts in the social network visualization, in preparation for the upcoming interviews. I searched the key word "social network analysis" and downloaded related articles or tip sheets from IRE website. I found I need to polish the interview questions a little bit to best utilize the strengths of different interviewees.

Week 12.

This week I was waiting for feedback from supervisor about my projects. I made lots of progress in conducting research and interviews. I finished conducting the interviews with Brant Houston, Jaimi Dowdell and Sarah Cohen. All of them have done pioneer work applying social network analysis in journalism. Brant Houston became interested in this area at the beginning of this century and loved to share his knowledge with his students. Jaimi did her master's thesis on social network analysis. But it is surprising to know that in recent years she didn't use this method a lot in the newsroom. She taught it at NICAR conference since 2004. Sarah Cohen didn't use it frequently either, mainly from the statistical perspective. Jaimi and Sarah both loved to use it as a reporting tool to serve the story.

I would need to make more research about the general for journalists to use social network analysis. It can be research, visualization, or both. And what are the obstacles for journalists to use it on a larger scale? It may be the deadline pressure, the technical difficulty or both. I could begin looking into the specific reasons in next couple of days.

Week 13.

This week, my coworker gave me some advice regarding the shape of sliders. My supervisor hasn't got back to me yet about specific revisions. I interviewed William Lyon from Neo4J. He is a Developer Relations Engineer at Neo4j. He also heads up the Neo4j Data Journalism Accelerator Program. It was a very successful interview since I learned a lot about the graph database and graph algorithms from talking with him. Some key concepts he introduced to me were kind of confusing, because I could not fully understand the relationship between the technique behind algorithms and the question it is going to solve. So I dived into the guide book released by Neo4j to study the details and categorization of different graph algorithms. William also introduced his experience leading the Neo4j Data Journalism Accelerator Program. He cooperated with NBC News in the investigation of Russian Twitter troll and gained a lot of journalistic insight about using social network analysis.

Mark Schifferli, the designer of Connected China introduced the technical details when he designed the network graphic. It was confirmed that Connected China was never a mere visualization that can do without journalists' effort. Irene Jay Liu flew from Hong Kong to Boston a lot to better communicate with the design team and discuss the findings and ideas. The design team needed to know more about the ecosystem of Chinese politics in order to further reflect the truth. Pierre Romera, the lead developer of the Influencers project introduced the process to produce that story. The investigation had been finished before the graphic was produced. In this project, the graphic totally served the story. Due the small size of dataset, they mainly focused on a more readable visual design and did not consider some complicated statistical algorithms

Week 14.

This is my last week at USA Today. I didn't have chance to talk about the project in detail with my supervisor. He said the draft of women votes would be handed over to the next intern and get published when the 100th anniversary of the passage of the 19th Amendment comes. I began transcribing interviews while contacting sources and continuing interviewing different reporters. Reg Chua opened up a new door in my research. He gave me lots of details and his own reflections regarding the workflow and drawback of Connected China. People tend to consider it as a very fancy project.

Reg considers it as a very good example of structure journalism. Sometimes, what is not new also has tremendous journalistic values. They worked very hard to compile a database of Chinese elites. But the problem is the audience would need some basic understanding of Chinese politics, so they could enjoy reading this graphic without paying lots of effort to learning new knowledge. Most of the readers may neither know nor care about the basic structure of Politburo Standing Committee. Reg came up with some more specific connections they could have looked into. Readers may have more interest in reading the story and the journalists may not need to paint the whole landscape.

2. research reports

Week 15.

I finished most of the transcriptions and began to think about possible angles of my professional analysis. Of course, it will revolve around my research questions or hypothesis. Some new questions began to emerge in my mind in the process of conducting interviews and browsing more research papers. Just as what my committee members talked about in the first meeting: is social network analysis really feasible? In what aspects? What could we do to figure out a flexible and powerful tool to meet the demands of journalists?

Consistent with the philosophy of computer-assisted reporting, social network analysis has to serve the story itself in every way possible. Lots of reporters don’t have statistical background and have to use ready-made tools and adapt to the current function of those tools. The graph database developers and data journalists could have more communication, especially in the aspect of developing new tools, education and training.

Week 16.

Jaimi Dowdell also inspired me to collect similar tools and resources together for journalists to pick. So I began to search around and compile a list of useful websites and software. Even if new tools keep coming up and replace old tools, I think it is still helpful to put together the endeavor of journalists and programmers to promote social network analysis in unravelling power relations and public discourses. I began to experiment with different social network tools I found online, such as Gephi, Neo4j. Most of them are very handy and efficient. But I couldn’t have a deeper understanding due to the lack of statistical background.

Week 17.

This week I spent the majority of my time transcribing interviews and finding relevant materials that my interviewees mentioned in our talk.

I also found lots of online tutorials are accessible for those starters who don’t have statistical knowledge or a rigorous academic plan. Science reporter Peter Aldhous taught NodeXL at NICAR frequently and shared tutorials on his Github account. Data reporter and researcher Jonathan Stray began to teach social network analysis to his students at Columbia Journalism University almost five years ago. Their tutorials gave me lots of help and inspirations when I encounter problems in my research.

Week 18.

This week I began to look into graph databases and started from Neo4j. Generally, they use three types of graph algorithms: centrality algorithms, community detection algorithms, pathfinding and traversal algorithms. The centrality algorithms include PageRank, Degree Centrality, Closeness Centrality and Betweenness Centrality. The pathfinding and traversal algorithms include Parallel Breadth-First search, Parallel Depth-First search, Single-Source shortest path, All-Pairs shortest path and Minimum Weight Spanning Tree. The community detection algorithms include Label Propagation, Strongly Connected Algorithms, Union-Find/Connected components, Louvain Modularity, Local Clustering coefficient, Triangle-Count coefficient.

They published a handy guidebook. I looked into those key concepts one by one and began to think about in what circumstances journalists may need them. To be honest, they have a much friendlier user interface than the graph algorithms handbook. We're not required to understand its logic completely before we jump into the software and play around.

Week 19.

This week I tried Gephi, a tool for exploring graph database. The user could manipulate the structures, shapes and colors to reveal hidden patterns. "The goal is to help data analysts to make hypothesis, intuitively discover patterns, isolate structure singularities or faults during data sourcing. It is a complementary tool to traditional statistics, as visual thinking with interactive interfaces is now recognized to facilitate reasoning. " Gephi explained its mission in this way, which is the core idea of Exploratory Data Analysis as well as investigative journalism.

The computer metrics of Gephi is similar with that of Neo4j: Betweenness Centrality, Closeness, Diameter, Clustering Coefficient, PageRank, Community detection (Modularity), Random generators, and Shortest path. But the visualization function of Gephi is more fancy and powerful. The Layout palette allows user to change layout settings and optimize for graph readability.

Week 20.

Reg Chua, editor of Connected China, recommended me to talk with Miguel Paz. I made more research about Miguel Paz and his work. Miguel Paz is the founder of Poderapedia, which reveals links among Chilean business and political leaders. This project is 10 years older than Connected China and Panama Papers. It's more similar with Connected China. Instead of chasing after the hot-button issues, they have done a great job in building a meta-data project that could provide context and background knowledge to journalists and scholars. And people could reuse it over and over again. There're a bunch of meta-data journalism projects, such as Theyrule.net. Who Runs Hong Kong. But unfortunately they were frozen, once project leaders jumped to the next job.

Week 21.

This week I tried the free version of NodeXL, an open-source template for Microsoft Excel to explore network graphs. It's a very handy tool to analyze the social media data but didn't have lots of options for us to customize the visualization. We can just import a network dataset in a worksheet and see the graph created automatically. I didn't try NodeXL Pro, which has access to social media network data streams, advanced network metrics, and text and sentiment analysis.

Marc Smith, one of the founders of Social Media Research Foundation and NodeXL, is also passionate about data journalism as well. He worked with Global Investigative Journalism Network to compile Top Ten data journalism project, using NodeXL to do a social network analysis of Twitter traffic with the #ddj hashtag.

Week 22.

I found an interesting phenomena that lots of projects that involved social network analysis were conducted by commercial consultants or academic researchers. It makes sense, because they may have more time and money, while journalists tend to be bound by tight deadlines and insufficient grants. That's also what the interviewees told me before. Data collection itself is already very demanding and time-consuming, let alone importing the data into social network analysis software and running graph algorithms.

Social media is inherently made up of network relationships and it's increasingly woven with the journalism and politics landscape and exerting a tremendous influence on the evolvement of public opinion in the United States. One of my research focuses is to explore the potential of social network analysis in untangling the social media discussion. I found some key researchers and scheduled interviews with them. I need to make more research, polish my questions and prepare well for the upcoming interview.

Week 23.

This week I spent lots of time reading papers and contacting sources, in an effort to explore how journalists or journalism researchers used social network analysis to map out the social media discourse. Some organizations, such as MIT Media Lab, Pew Research Center, Graphika, did a great job showing how social media accounts are connected to each other through social relationships embedded in the platform--in the case of Twitter, through patterns of followership or retweets. I looked into how they created their social network maps and what kind of insights they were able to offer on a given topic, unravel the relationship of different interest groups and advance civic engagement.

Week 24.

Miguel Paz looked back on the history of Poderapedia in our interview. The social network data they compiled and fact-checked is really a huge help to journalists and academic researchers. Unfortunately that website stopped updating after Miguel Paz moved to New York. The funding from Knight Foundation was running out too.

Marc Smith helped deepen my understanding about social media network. He began digging into sociology since the Utopian community and Internet took off several decades ago. He had a huge passion about visualizing the online community and co-founded the Connected Action and Social Media Research Foundation. He hopes that more and more journalists could utilize the tool of NodeXL and take an aerial photo of social media. In this way, journalists could find stories from the shape and structure of different clusters. I looked into some academic papers he published. And I got fascinated by the clear-cut and easy step for NodeXL to find emergent clusters on Twitter. He also categorized six types of clusters and clarified different circumstances in which journalists may need them to explore the public discourse.

Unfortunately, very few journalists have adopted his methods when analyzing the landscape of social media discussion.

Week 25.

This week I was mainly working on the interaction between journalistic judgment and graph algorithms this week. It is a very interesting topic that percolates through the history of computer-assisted reporting. The case study of Connected China contributed to my analysis. Reg and Mark inspired me to dig deeper into this topic. It would be great to have the opinion of project leader as well. Therefore I reached out to Irene Jay Liu several times, but she didn’t respond.

In general, Connected China is a valuable project that provides context and background knowledge to lots of readers who’re interested in the Chinese elite politics. But it didn’t have a concrete story that could serve as the bone of the whole project. There’re too many threads, or too may sceneries. You may wander around during the journey. Readers may spend half an hour on this project. For those people who don’t have an inkling of Chinese politics, Connected China asks them to learn lots of new material at one time.

CHAPTER THREE: EVALUATION

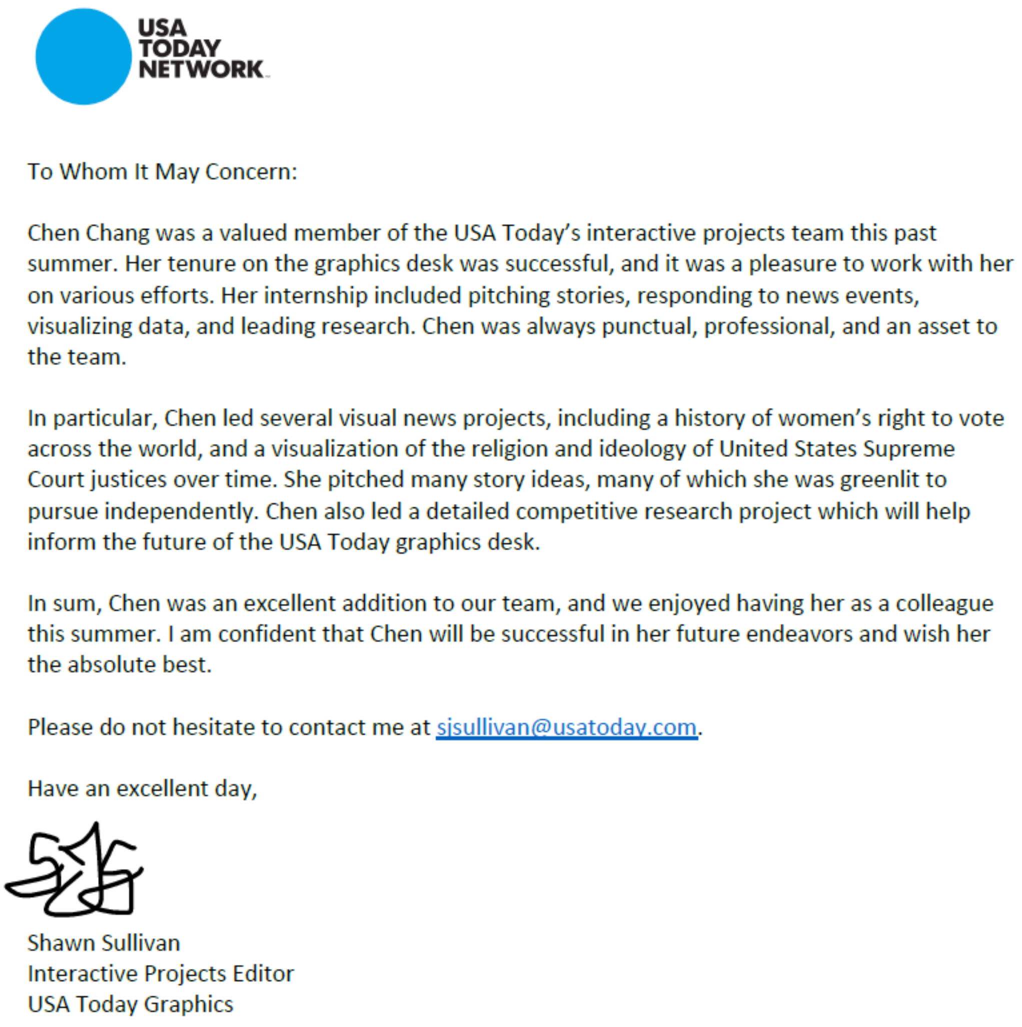
Self-evaluation

The graphic team has a very specific role in the newsroom. They are mainly in charge of creating graphics for social media and breaking news every day. They complete day-to-day tasks with a strong work ethic and meticulous standards, which has taught me a lot about the duty and attitude of a good designer in the newsroom. In terms of projects, such as a long-form investigative story, the graphic team sometimes cooperates with news reporters and helps with illustrations and web design. But they hardly start a project themselves and focus on it independently. Considering the infrequent interaction between graphics team and reporting team, I tried to come up with journalistic ideas and work on it independently. I pitched some story ideas to my supervisor and made progress in collecting data, cleaning data, making GIF and interactive graphics with JavaScript. In this process, I've learned to find a simple and engaging way to visualize a story and take users into consideration in web design. For example, one project is about how women's voting rights spread all over the world in history. I looked through some textbooks and learned to combine the chronological data and geographic data in D3. USA Today doesn't plan to publish it until probably 2020, which marks the 100th anniversary of the passage of the 19th Amendment, guaranteeing women's constitutional right to vote.

I could have taken more initiative and communicated more with my supervisor. It was a pity that I didn't get to take part in a project led by both reporters and designers. One big project Maternal Harm was being wrapped up when I first came to the office. Another project Toxic City was about to start, at the end of my internship. But if I had reached out to other branches in the newsroom, I may have had some more chance to help with data analysis and reporting. In general, I spent much time on learning D3 and debugging skills, rather than having a big picture about the goal of this internship and tried to grasp more journalistic techniques. Other candidates may have been able to contribute more to the team if they had been given this opportunity.

Evaluation from USA Today

Below is the feedback from my supervisor Shawn Sullivan:



USA TODAY NETWORK

To Whom It May Concern:

Chen Chang was a valued member of the USA Today's interactive projects team this past summer. Her tenure on the graphics desk was successful, and it was a pleasure to work with her on various efforts. Her internship included pitching stories, responding to news events, visualizing data, and leading research. Chen was always punctual, professional, and an asset to the team.

In particular, Chen led several visual news projects, including a history of women's right to vote across the world, and a visualization of the religion and ideology of United States Supreme Court justices over time. She pitched many story ideas, many of which she was greenlit to pursue independently. Chen also led a detailed competitive research project which will help inform the future of the USA Today graphics desk.

In sum, Chen was an excellent addition to our team, and we enjoyed having her as a colleague this summer. I am confident that Chen will be successful in her future endeavors and wish her the absolute best.

Please do not hesitate to contact me at sjsullivan@usatoday.com.

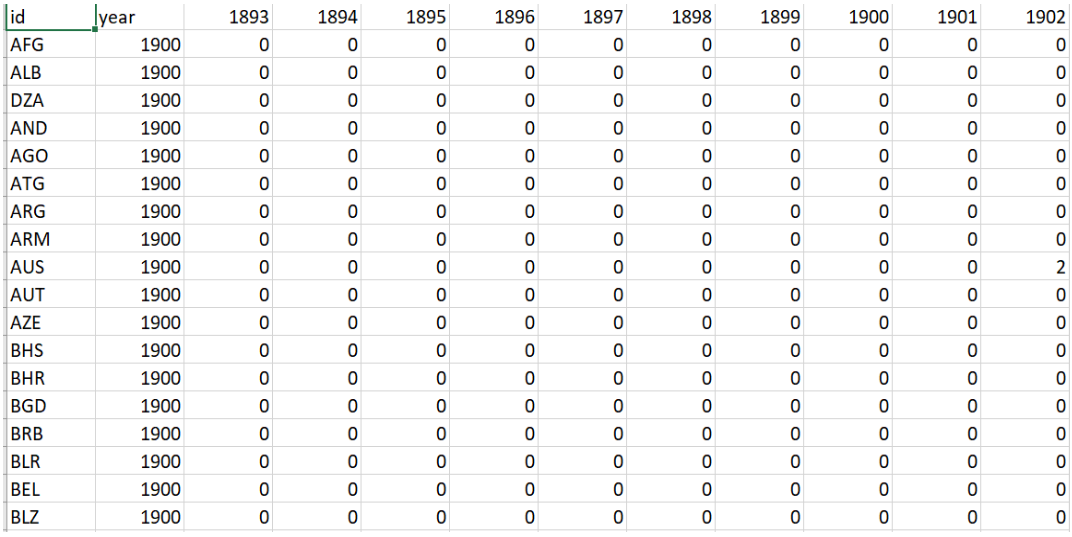
Have an excellent day,

Shawn Sullivan
Interactive Projects Editor
USA Today Graphics

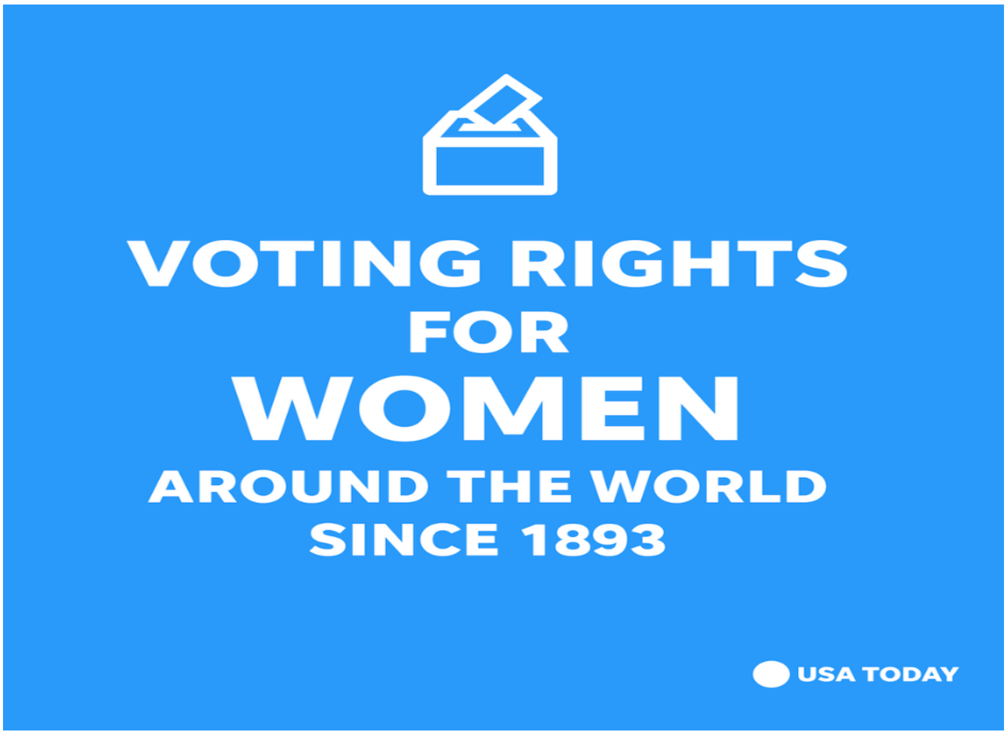
CHAPTER 4: PHYSICAL EVIDENCE

Women’s suffrage project

This project displays how women’s suffrage spread all over the world from 1893 to 2011. The static graphics are designed for social media promotion. The interactive map allows readers to control the year and see the historical trend in a dynamic way.

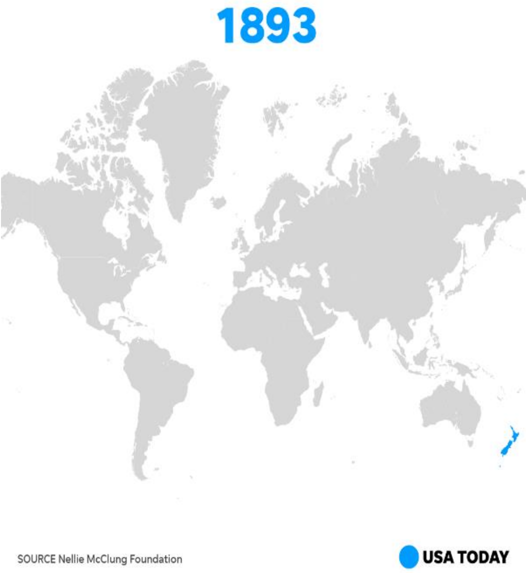


id	year	1893	1894	1895	1896	1897	1898	1899	1900	1901	1902
AFG	1900	0	0	0	0	0	0	0	0	0	0
ALB	1900	0	0	0	0	0	0	0	0	0	0
DZA	1900	0	0	0	0	0	0	0	0	0	0
AND	1900	0	0	0	0	0	0	0	0	0	0
AGO	1900	0	0	0	0	0	0	0	0	0	0
ATG	1900	0	0	0	0	0	0	0	0	0	0
ARG	1900	0	0	0	0	0	0	0	0	0	0
ARM	1900	0	0	0	0	0	0	0	0	0	0
AUS	1900	0	0	0	0	0	0	0	0	0	2
AUT	1900	0	0	0	0	0	0	0	0	0	0
AZE	1900	0	0	0	0	0	0	0	0	0	0
BHS	1900	0	0	0	0	0	0	0	0	0	0
BHR	1900	0	0	0	0	0	0	0	0	0	0
BGD	1900	0	0	0	0	0	0	0	0	0	0
BRB	1900	0	0	0	0	0	0	0	0	0	0
BLR	1900	0	0	0	0	0	0	0	0	0	0
BEL	1900	0	0	0	0	0	0	0	0	0	0
BLZ	1900	0	0	0	0	0	0	0	0	0	0



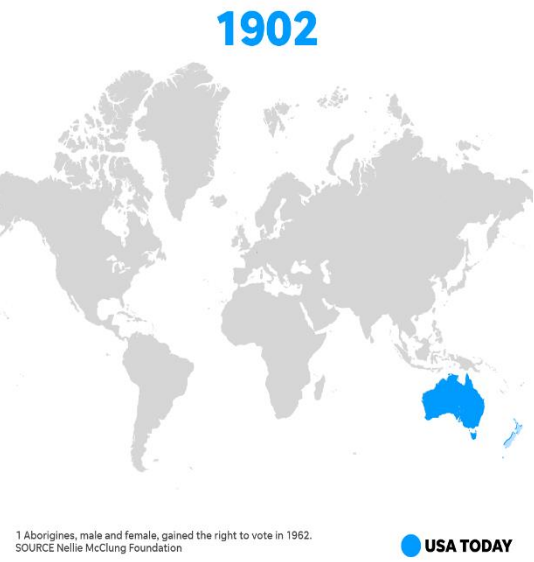
VOTING RIGHTS
FOR
WOMEN
AROUND THE WORLD
SINCE 1893

USA TODAY



SOURCE Nellie McClung Foundation

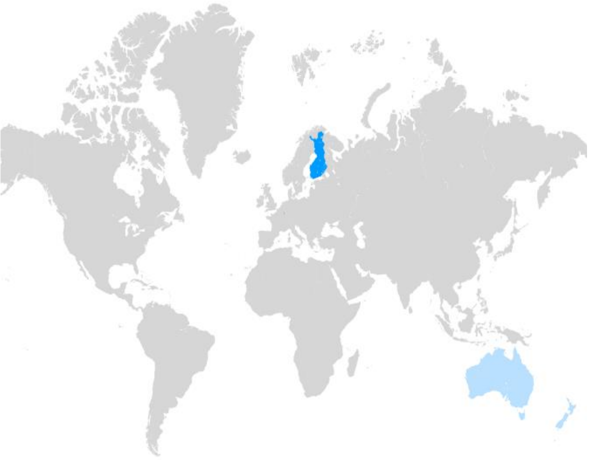
USA TODAY



1 Aborigines, male and female, gained the right to vote in 1962.
SOURCE Nellie McClung Foundation

USA TODAY

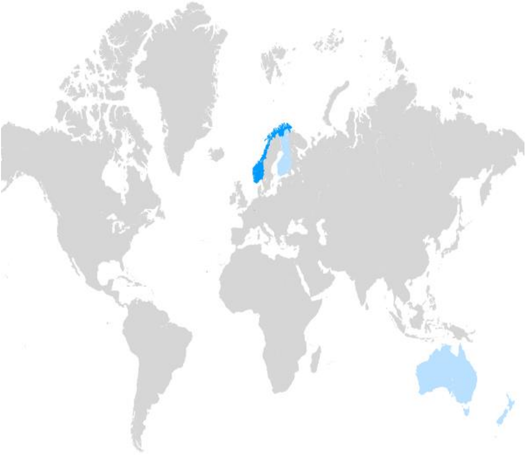
1906



SOURCE Nellie McClung Foundation

USA TODAY

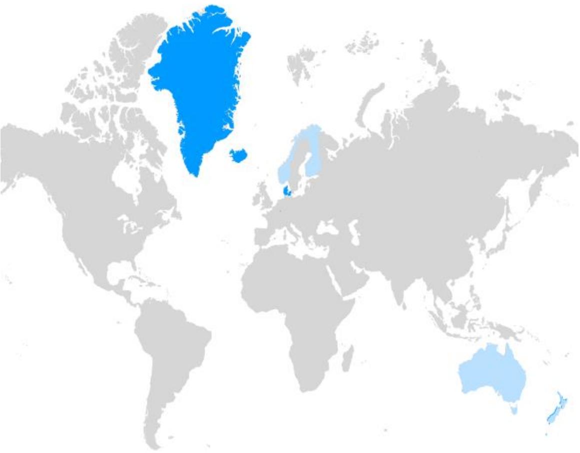
1913



SOURCE Nellie McClung Foundation
Inter-Parliamentary Union
Women Suffrage and Beyond/University of British Columbia

USA TODAY

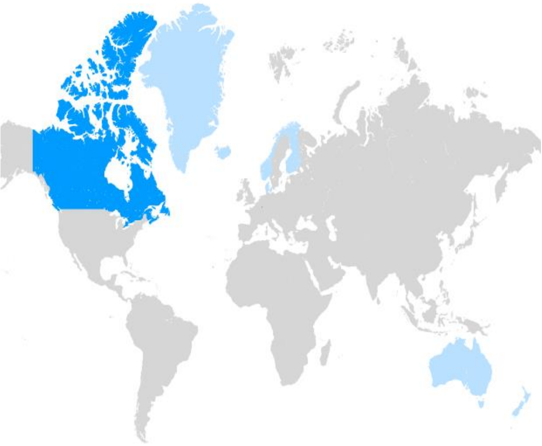
1915



NOTE: Greenland is part of the Kingdom of Denmark.
SOURCE Nellie McClung Foundation
Inter-Parliamentary Union
Women Suffrage and Beyond/University of British Columbia

USA TODAY

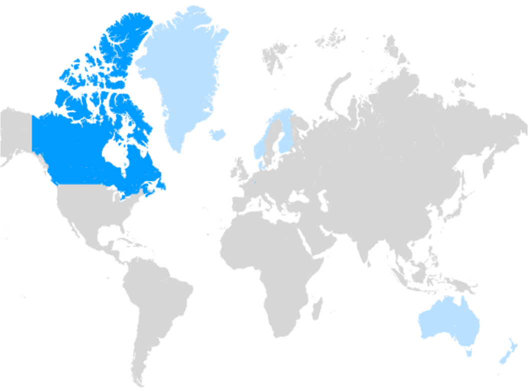
1917



2 Canadian First Nation, male and female, did not win the vote until 1960.
SOURCE Nellie McClung Foundation
Inter-Parliamentary Union
Women Suffrage and Beyond/University of British Columbia

USA TODAY

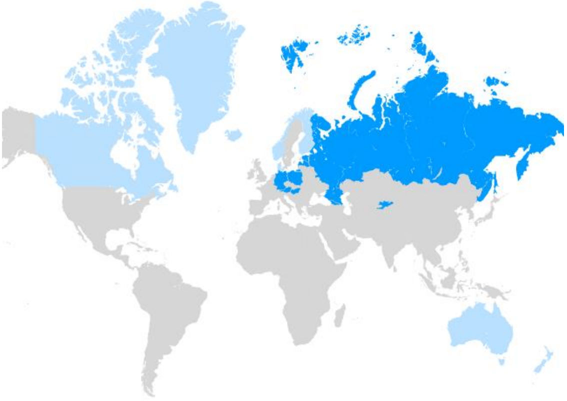
1917



2 Canadian First Nation, male and female, did not win the vote until 1960.
SOURCE Nellie McClung Foundation
Inter-Parliamentary Union
Women Suffrage and Beyond/University of British Columbia

USA TODAY

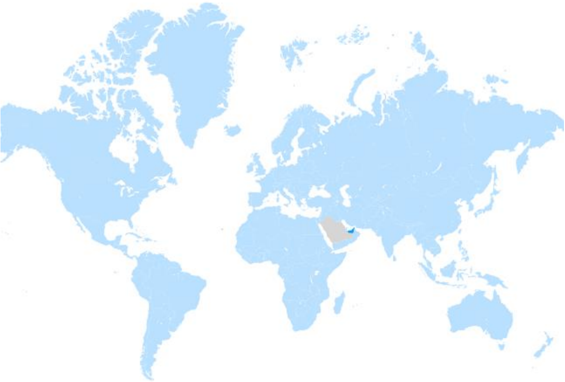
1918



SOURCE Nellie McClung Foundation
Inter-Parliamentary Union
Women Suffrage and Beyond/University of British Columbia

USA TODAY

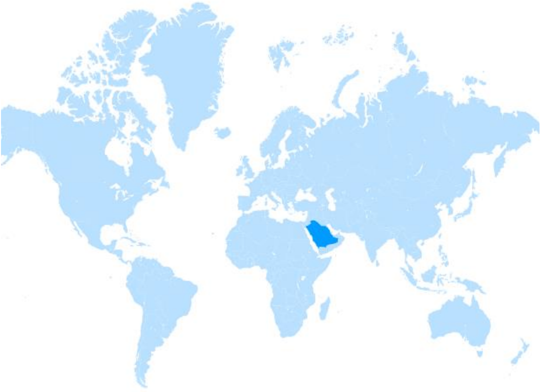
2006



SOURCE Nellie McClung Foundation
Inter-Parliamentary Union
Women Suffrage and Beyond/University of British Columbia

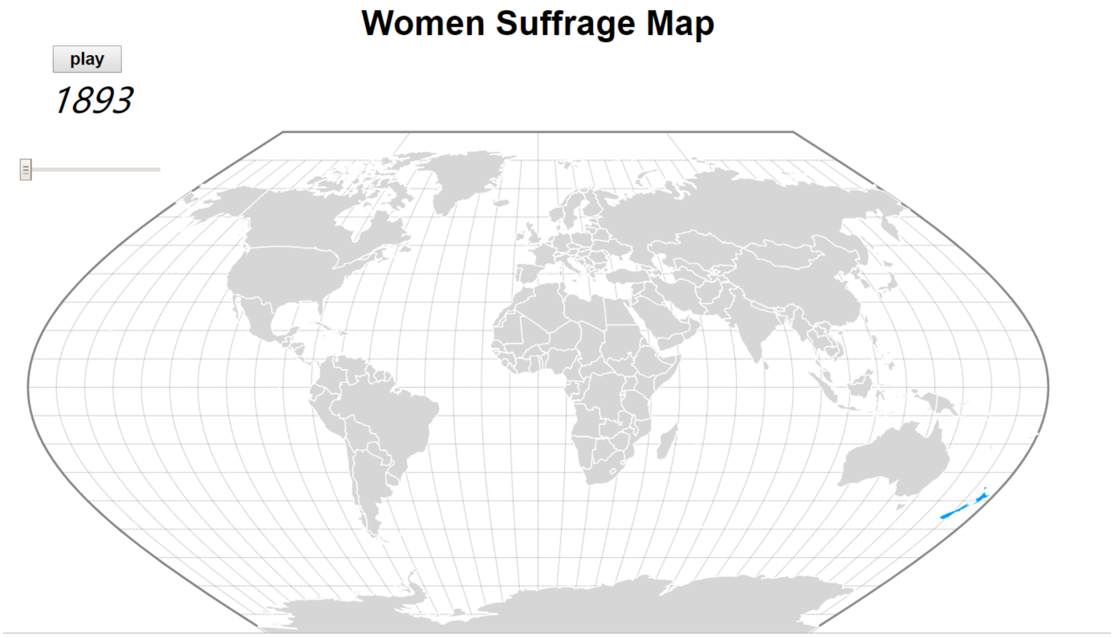
USA TODAY

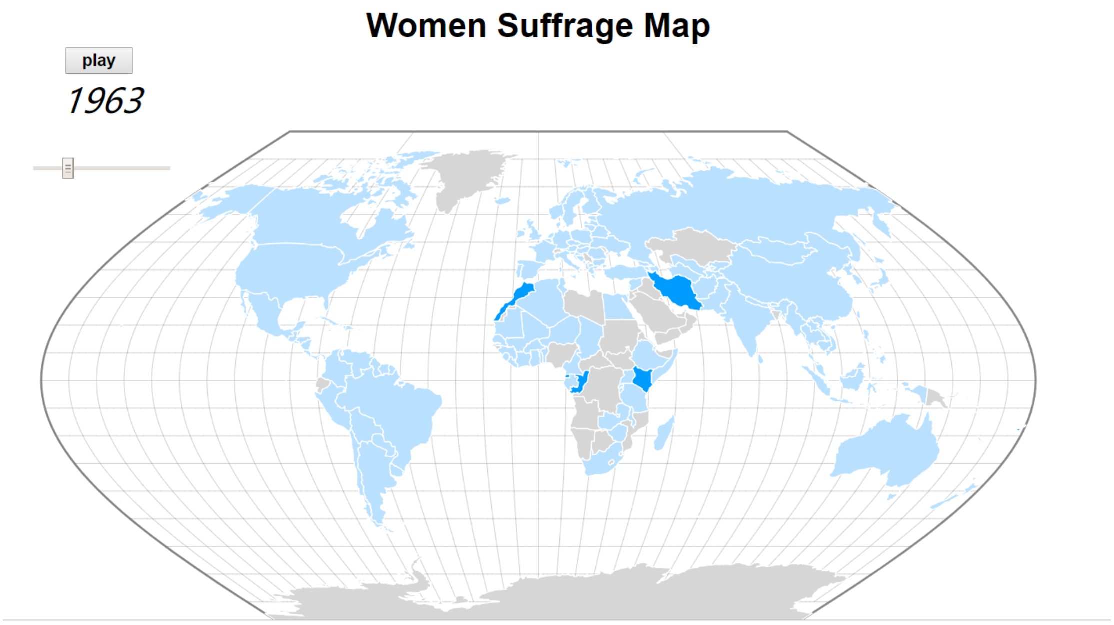
2011

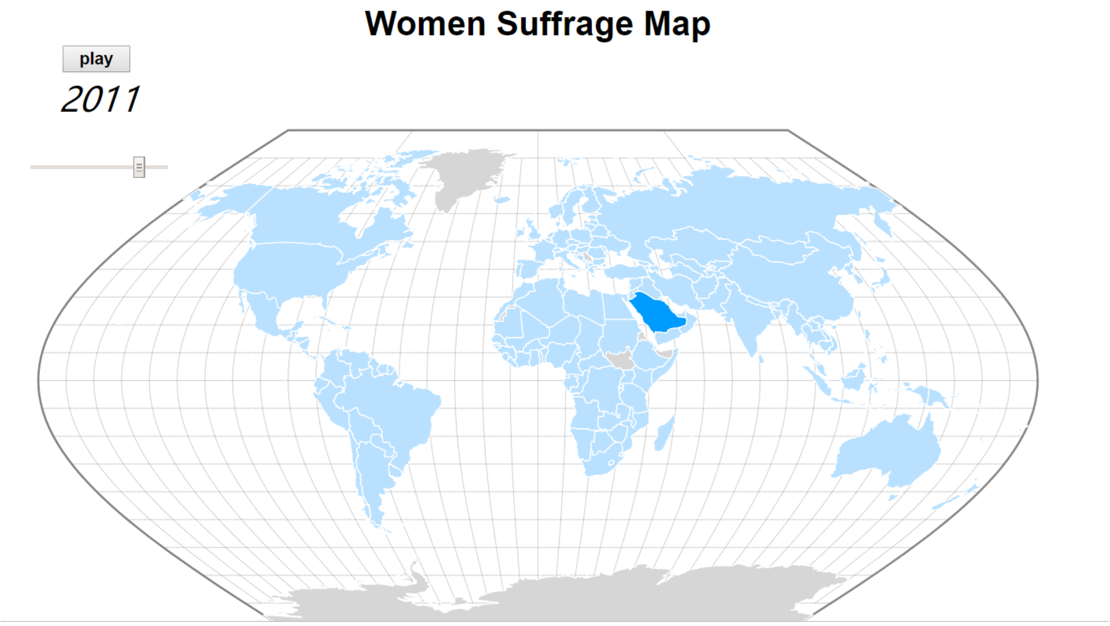


SOURCE Nellie McClung Foundation
Inter-Parliamentary Union
Women Suffrage and Beyond/University of British Columbia

USA TODAY

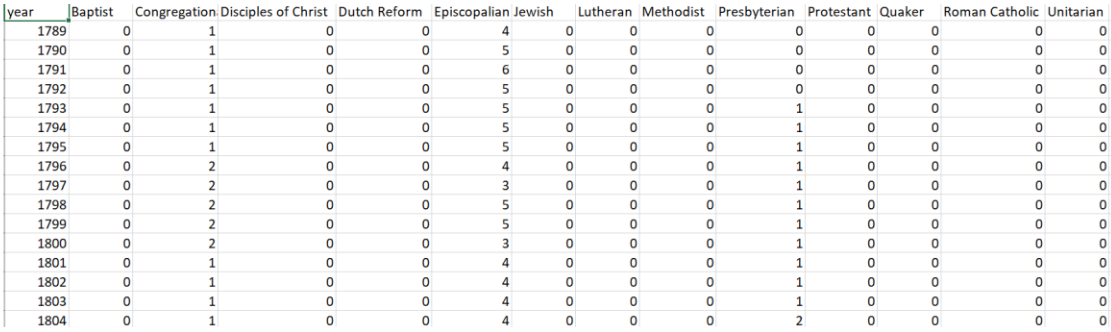

Women Suffrage Map
play
1893


Women Suffrage Map
play
1963

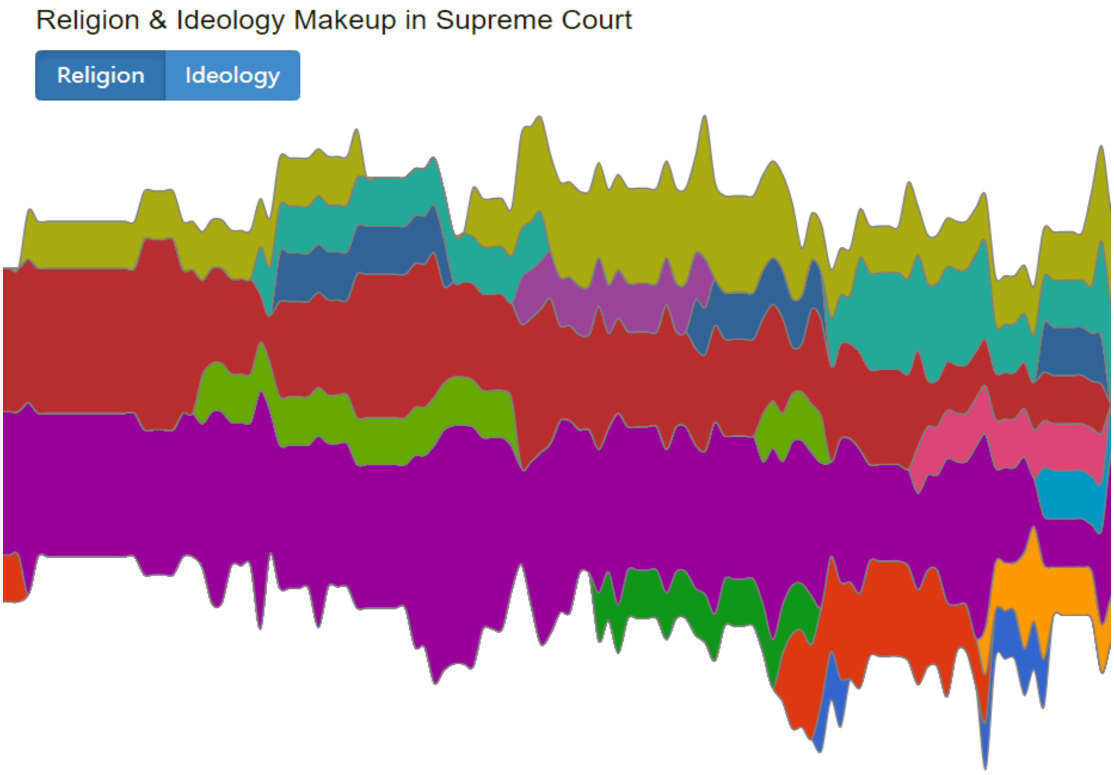


Supreme court project

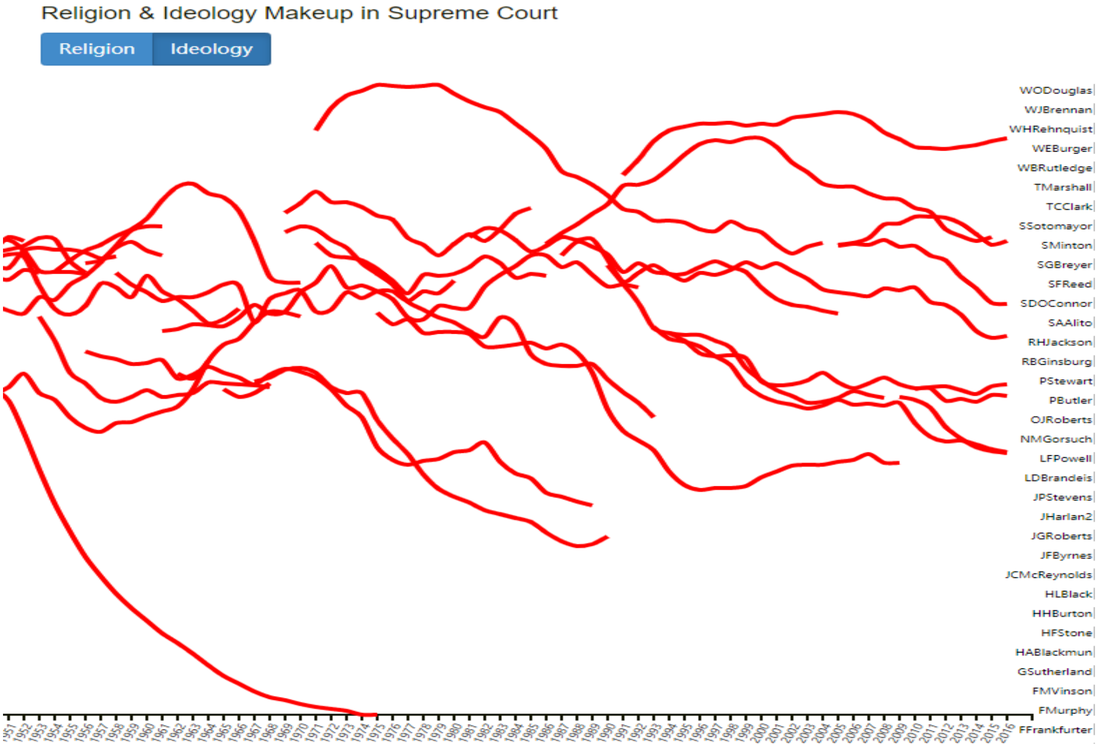
I’ve collected research data from scholars Lee Epstein, Andrew D. Martin, and Kevin Quinn. They place judges on an ideological spectrum called the “Judicial Common Space." Conservative justices receive scores from 0 to 1, liberal justices from –1 to 0. In this way, we’re able to compare the ideological leans and religious backgrounds of each justice in U.S. Supreme Court.



year	Baptist	Congregation	Disciples of Christ	Dutch Reform	Episcopalian	Jewish	Lutheran	Methodist	Presbyterian	Protestant	Quaker	Roman Catholic	Unitarian
1789	0	1	0	0	4	0	0	0	0	0	0	0	0
1790	0	1	0	0	5	0	0	0	0	0	0	0	0
1791	0	1	0	0	6	0	0	0	0	0	0	0	0
1792	0	1	0	0	5	0	0	0	0	0	0	0	0
1793	0	1	0	0	5	0	0	0	1	0	0	0	0
1794	0	1	0	0	5	0	0	0	1	0	0	0	0
1795	0	1	0	0	5	0	0	0	1	0	0	0	0
1796	0	2	0	0	4	0	0	0	1	0	0	0	0
1797	0	2	0	0	3	0	0	0	1	0	0	0	0
1798	0	2	0	0	5	0	0	0	1	0	0	0	0
1799	0	2	0	0	5	0	0	0	1	0	0	0	0
1800	0	2	0	0	3	0	0	0	1	0	0	0	0
1801	0	1	0	0	4	0	0	0	1	0	0	0	0
1802	0	1	0	0	4	0	0	0	1	0	0	0	0
1803	0	1	0	0	4	0	0	0	1	0	0	0	0
1804	0	1	0	0	4	0	0	0	2	0	0	0	0



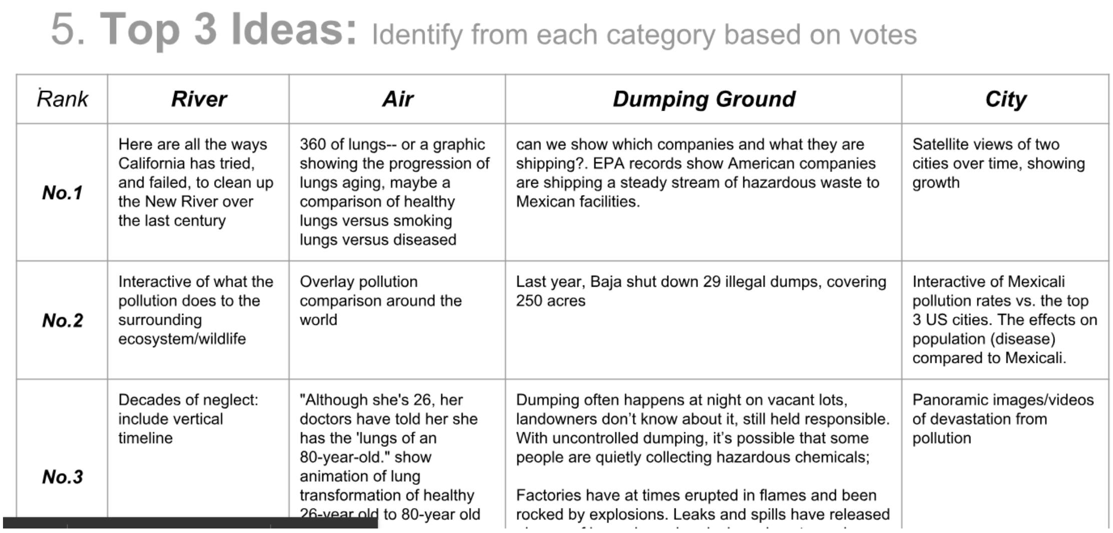
Religion & Ideology Makeup in Supreme Court



Religion & Ideology Makeup in Supreme Court

Toxic city project

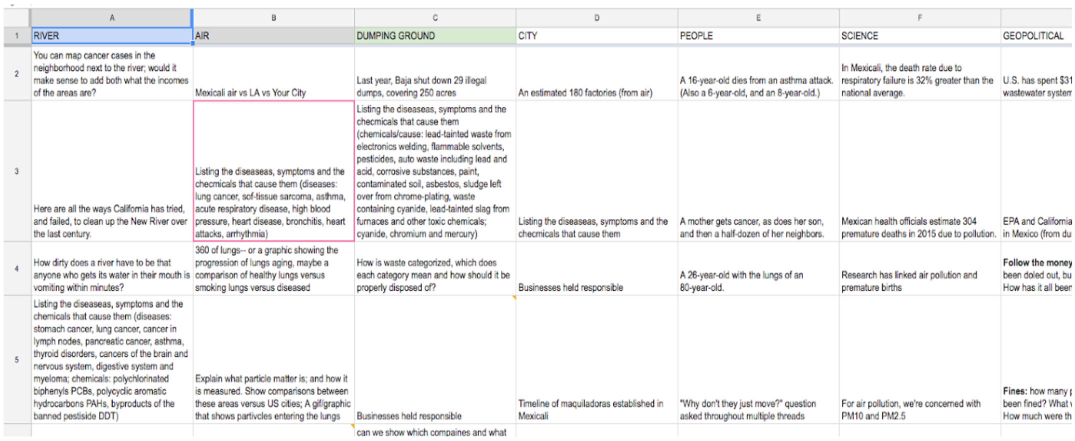
The USA Today Storytelling Studio worked on a project about the air, river and soil pollution along the Mexico–United States border. I attended the brainstorming meeting and made a slideshow to summarize some possible story angles with another intern.



5. **Top 3 Ideas:** Identify from each category based on votes

Rank	River	Air	Dumping Ground	City
No.1	Here are all the ways California has tried, and failed, to clean up the New River over the last century	360 of lungs-- or a graphic showing the progression of lungs aging, maybe a comparison of healthy lungs versus smoking lungs versus diseased	can we show which companies and what they are shipping?. EPA records show American companies are shipping a steady stream of hazardous waste to Mexican facilities.	Satellite views of two cities over time, showing growth
No.2	Interactive of what the pollution does to the surrounding ecosystem/wildlife	Overlay pollution comparison around the world	Last year, Baja shut down 29 illegal dumps, covering 250 acres	Interactive of Mexicali pollution rates vs. the top 3 US cities. The effects on population (disease) compared to Mexicali.
No.3	Decades of neglect: include vertical timeline	"Although she's 26, her doctors have told her she has the 'lungs of an 80-year-old." show animation of lung transformation of healthy 26-year old to 80-year old	Dumping often happens at night on vacant lots, landowners don't know about it, still held responsible. With uncontrolled dumping, it's possible that some people are quietly collecting hazardous chemicals; Factories have at times erupted in flames and been rocked by explosions. Leaks and spills have released	Panoramic images/videos of devastation from pollution

3. **Key Points:** Tease-out key ideas and topics under each category



	A	B	C	D	E	F	
1	RIVER	AIR	DUMPING GROUND	CITY	PEOPLE	SCIENCE	GEOPOLITICAL
2	You can map cancer cases in the neighborhood next to the river; would it make sense to add both what the incomes of the areas are?	Mexicali air vs LA vs Your City	Last year, Baja shut down 29 illegal dumps, covering 250 acres	An estimated 180 factories (from air)	A 16-year-old dies from an asthma attack. (Also a 6-year-old, and an 8-year-old.)	In Mexicali, the death rate due to respiratory failure is 32% greater than the national average.	U.S. has spent $31 wastewater system
3	Here are all the ways California has tried, and failed, to clean up the New River over the last century.	Listing the diseaseas, symptoms and the checmicals that cause them (diseases: lung cancer, sof-tissue sarcoma, asthma, acute respiratory disease, high blood pressure, heart disease, bronchitis, heart attacks, arrhythmia)	Listing the diseaseas, symptoms and the checmicals that cause them (chemicals/cause: lead-tainted waste from electronics welding, flammable solvents, pesticides, auto waste including lead and acid, corrosive substances, paint, contaminated soil, asbestos, sludge left over from chrome-plating, waste containing cyanide, lead-tainted slag from furnaces and other toxic chemicals; cyanide, chromium and mercury)	Listing the diseaseas, symptoms and the checmicals that cause them	A mother gets cancer, as does her son, and then a half-dozen of her neighbors.	Mexican health officials estimate 304 premature deaths in 2015 due to pollution.	EPA and California in Mexico (from du
4	How dirty does a river have to be that anyone who gets its water in their mouth is vomiting within minutes?	360 of lungs-- or a graphic showing the progression of lungs aging, maybe a comparison of healthy lungs versus smoking lungs versus diseased	How is waste categorized, which does each category mean and how should it be properly disposed of?	Businesses held responsible	A 26-year-old with the lungs of an 80-year-old.	Research has linked air pollution and premature births	**Follow the money** been doled out, bu How has it all been
5	Listing the diseaseas, symptoms and the chemicals that cause them (diseases: stomach cancer, lung cancer, cancer in lymph nodes, pancreatic cancer, asthma, thyroid disorders, cancers of the brain and nervous system, digestive system and myeloma; chemicals: polychlorinated biphenyls PCBs, polycyclic aromatic hydrocarbons PAHs, byproducts of the banned pestiside DDT)	Explain what particle matter is; and how it is measured. Show comparisons between these areas versus US cities; A gif/graphic that shows partivcles entering the lungs	Businesses held responsible	Timeline of maquiladoras established in Mexicali	"Why don't they just move?" question asked throughout multiple threads	For air pollution, we're concerned with PM10 and PM2.5	**Fines:** how many p been fined? What v How much were th
			can we show which compaines and what				

6. **Top ideas:** Identify from each category based on votes

- **River:** Here are all the ways California has tried, and failed, to clean up the New River over the last century
- **Air**: **360 of lungs**-- or a graphic showing the progression of lungs aging, maybe a comparison of healthy lungs versus smoking lungs versus diseased
- **Dumping Ground:** Can we show which companies and what they are shipping?. EPA records show American companies are shipping a steady stream of hazardous waste to Mexican facilities
- **City: Satellite views** of two cities over time, showing growth
- **People:** Use audio to show what it like to have an asthma attack. See: Salton Sea
- **Science:** For air pollution, we're concerned with PM10 and PM2.5
- **Geopolitical:** There's a flow, back and forth across the border: Companies go into Mexicali, but so does US waste to be dumped. Consumer goods come out, but so does pollution and illegal immigration

Besides, I cooperated with another intern to conduct research on the graphics desks in around ten news organizations. I looked into their duties, structures, strengths and summarized key strategies that can be applied in the graphics team at USA Today. It is an online spreadsheet which I don't have access to right now.

CHAPTER FIVE: PROFESSIONAL ANALYSIS

Social network analysis allows journalists to take an aerial picture of the social networks, rather than taking a snapshot of a small group or certain individuals. It enables journalists to discover the key players, hidden ties, clusters, structures and patterns of the social networks, especially when they analyze complicated power relations in investigative journalism. However, social network analysis has not taken off since it was first brought to the journalism industry two decades ago. This article mainly digs into the potential of social network analysis as a powerful reporting tool, as well as its shortcomings that prevents journalists from applying it on a larger scale.

Three major findings emerged from my research and interviews regarding Connected China by Reuters and The Influencers by the International Consortium of Investigative Journalists. First, when it comes to the workflow, data journalists start from data collection, use graph algorithms or graph database tools to test hypotheses, and then produce a network visualization to display the results in a vivid and clear way. Second, the social network analysis is a useful reporting tool. Graph algorithms can help journalists make breakthroughs in investigations. They can provide a broad picture of structures and patterns across the network efficiently and automatically. Third, several limits of social network analysis may be obstacles to its widespread application. These limits include time-consuming data collection, unpredictable outcomes, unsustainable databases, and complicated interaction between the algorithmic analysis and editorial judgment.

Process: Data collection, Network Analysis and Visualization

The whole process usually involves three steps. First, data journalists need to collect data from a variety of sources. Second, data journalists use graph algorithms or graph database tools to test hypotheses and shed light on valuable information hidden in a connected network. Third, data journalists and designers produce a network visualization to display the whole structure or emphasize key players.

Data collection usually requires journalists to look into a variety of public records. Journalists and researchers working on Connected China spent several months collecting data from official websites, news archives, scholarly publications and a variety of data sources and built their own database to track the social network of Chinese officials. Connected China includes tens of thousands of entities and 30,000 relationships that were identified and typed into the database by journalists and researchers. This database is only accessible to Reuters reporters. Journalists working on The Influencers, in contrast, didn't encounter considerable difficulty in data collection. They used what the ICIJ had already gotten for The Paradise Papers, a global investigation based on 13.4 million leaked files from leading offshore law firm Appleby, trust company Asiaciti, and from company registries in 19 tax haven jurisdictions.

In her master's professional project about social network analysis in 2004, Jaimi Dowdell, now a data reporter at Reuters, compiled a list of useful public records or website sources that journalists could use to build such databases: local government websites, secretary of state corporate filings, U.S. Securities and Exchange Commission filings, Form 990s, newspaper archives, corporate websites, property records and court records. She categorized the data sources into local government power, campaign contributions, crime, public health, contracts and bids.

After collecting the data, some journalists would import it to graph database tools to conduct network analysis. Due to the technical difficulty and tight deadline, a social network map is used more frequently than algorithm analysis.

Connected China used algorithms to scale the importance of officials' political influence and indicate the possible affinity between different officials. The Influencers didn't use algorithm analysis, for the size of the data set was relatively small. The Influencers only included 13 players who are related to the Trump administration. Reporters singled out those stakeholders from the graph database of the Paradise Papers and began to conduct an investigation.

In contrast to the whole network or landscape painted in Connected China, the ego-network, which means a social circle around a certain actor within the network, allows journalists to develop narrower angles and have more interaction with editors. For example, the Influencers project presented an ego-network of Donald Trump following a complete story in the Paradise Papers. At first, Sasha Chavkin and Spencer Woodman finished researching and writing about Trump's allies who appeared in the Paradise Papers database. Then, Pierre Romera was in charge of developing a network visualization.

"I proposed this visualization where you have this network that is displayed step-by-step instead of the whole network at the same time. For each person, we tried to display the stories within steps." Romera said, "In most of the cases, we were able to have the entire network around an individual, select some part of the network, to zoom in and be more focused on the elements."

The Influencers project embeds the narrative text into an interactive graphic. ICIJ created a linear slideshow that explains the structure of different regions of the graph. We can select different nodes, find related people and read their profiles accordingly.

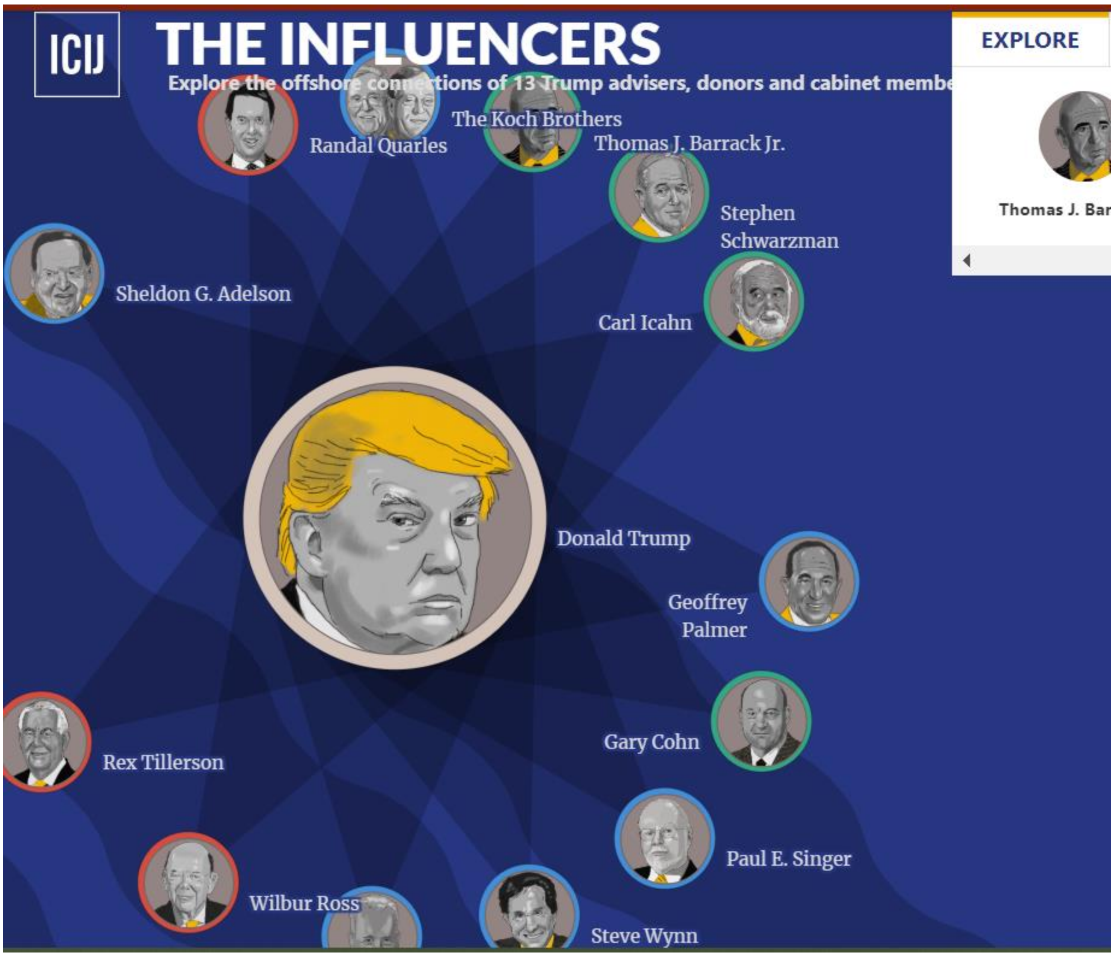
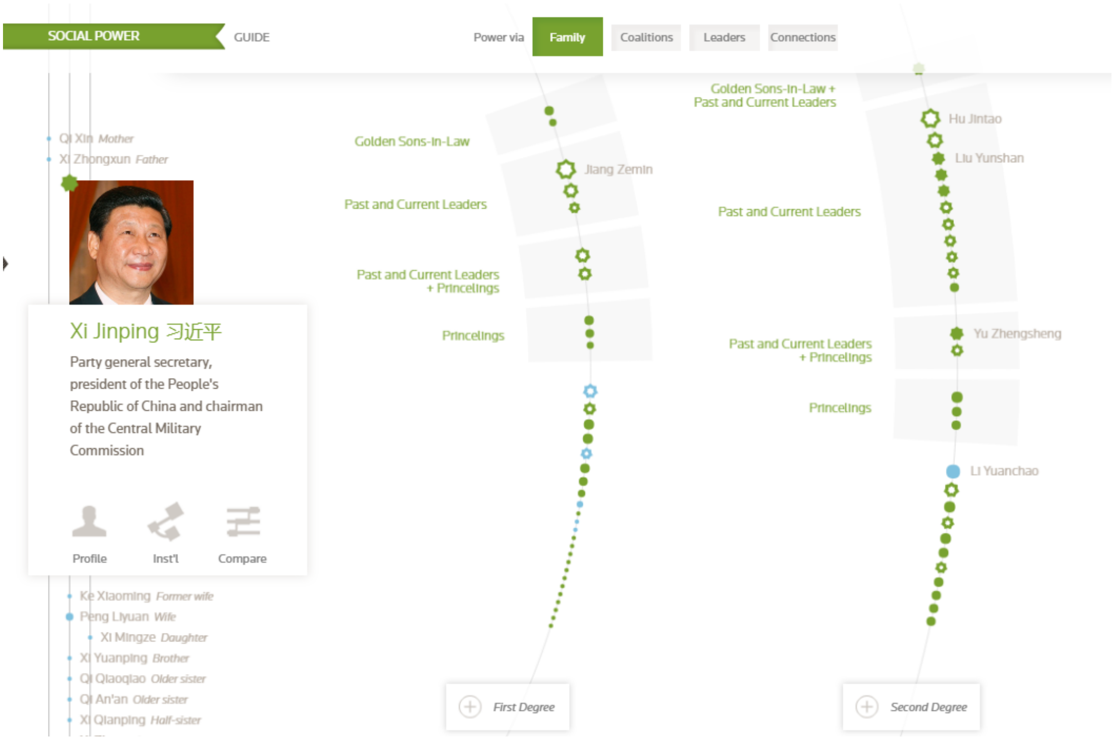


Figure 1. A screenshot from “The Influencers”, showing Trump’s social connections.

The Connected China project has a much more complicated narrative and went through lots of trial and error in its visual layout. At first, Reuters journalists came up with a simple model that allowed the user to center the network on a person and choose how many degrees of separation to show. Instead of showing a complex hairball of connections that may be visually disruptive, they settled on two degrees of connections and tried to emphasize the family connections. They put those people connected with an ego’s family member in the first degree. For example, Xi Jinping is the ego in the network as below. Below him are his family members, marked as blue dots. They also used that importance score to size an icon for an official to show the different political influence.



Party general secretary, president of the People's Republic of China and chairman of the Central Military Commission

Figure 2. A screenshot from “Connected China”, showing Xi Jinping’s social connections.

Jonathan Stray, a computational journalist and instructor at Columbia Journalism School, reviewed the existing uses of network analysis in journalism by analyzing a set of 34 completed stories since the late 1980s. He found five general attributes of network stories: story visualizations, scraping, reporting visualizations, algorithm and graph databases. Network visualizations are much more popular than graph algorithms.

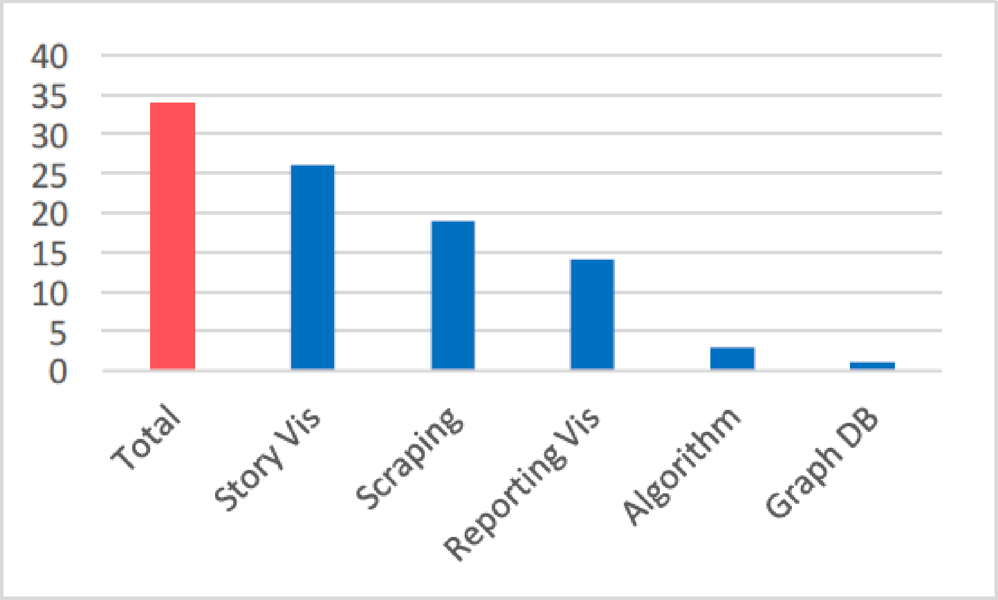


Figure 3. Attributes of the 34 network analysis stories collected by Jonathan Stray in Network Analysis in Journalism: Practices and Possibilities.

Jonathan Stray further came up with a sketch that illustrated the process to use social network analysis in newsrooms.

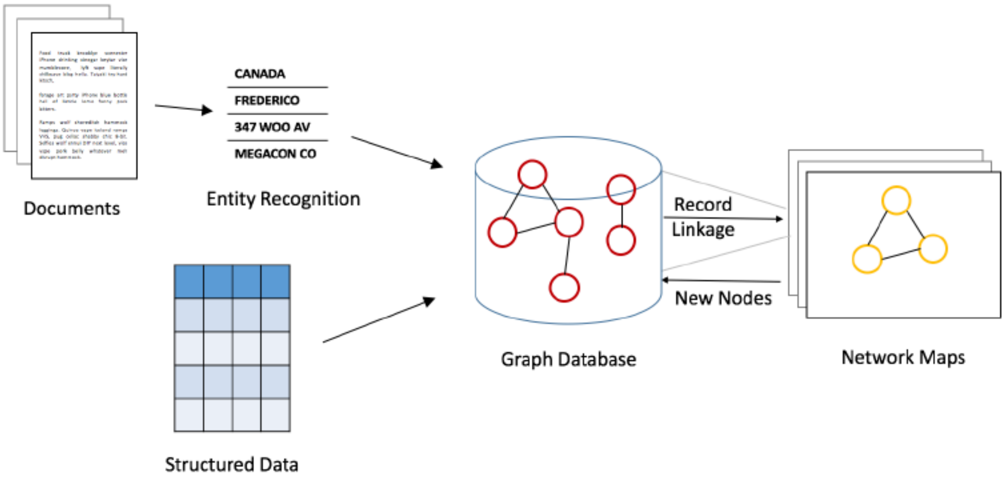


Figure 4. The proposed design for an integrated system for network analysis in investigative journalism by Jonathan Stray.

Reporting Tool for Investigation: Potential of Graph Algorithms

Based on my research interviews, while Connected China and The Influencers haven’t fully tapped into the potential of graph algorithms, some algorithmic techniques, such as centrality algorithms, community detection algorithms and path-finding algorithms are useful methods to identify key players and possible social connections.

The centrality algorithms are mostly used in data journalism to determine the influence and importance of distinct nodes in the network. The community detection algorithms, also known as clustering and partitioning algorithms, can be used to detect co-consumption networks, communication networks, geographical communities in data journalism. The path-finding algorithms can help journalists find two nodes that are connected to one another and the shortest path between two nodes.

William Lyon, a Developer Relations Engineer at Neo4j, heads up the company’s Data Journalism Accelerator Program to help data journalists investigate social networks and utilize graph databases. He especially emphasized the value of graph algorithms in the analysis of large databases.

For example, William cooperated with journalists from NBC News to investigate how Russian operatives tried to influence the 2016 U.S. presidential election via Twitter accounts and other social media platforms. They applied the community detection algorithm to the retweet network and found that that the graph partitions into three distinct clusters or communities. Then they ran the PageRank algorithm, which is an centrality algorithm that evaluates the quality and quantity of links to a webpage, to identify the most influential accounts within each cluster and detect and understand the patterns of behavior reflected by those connections.

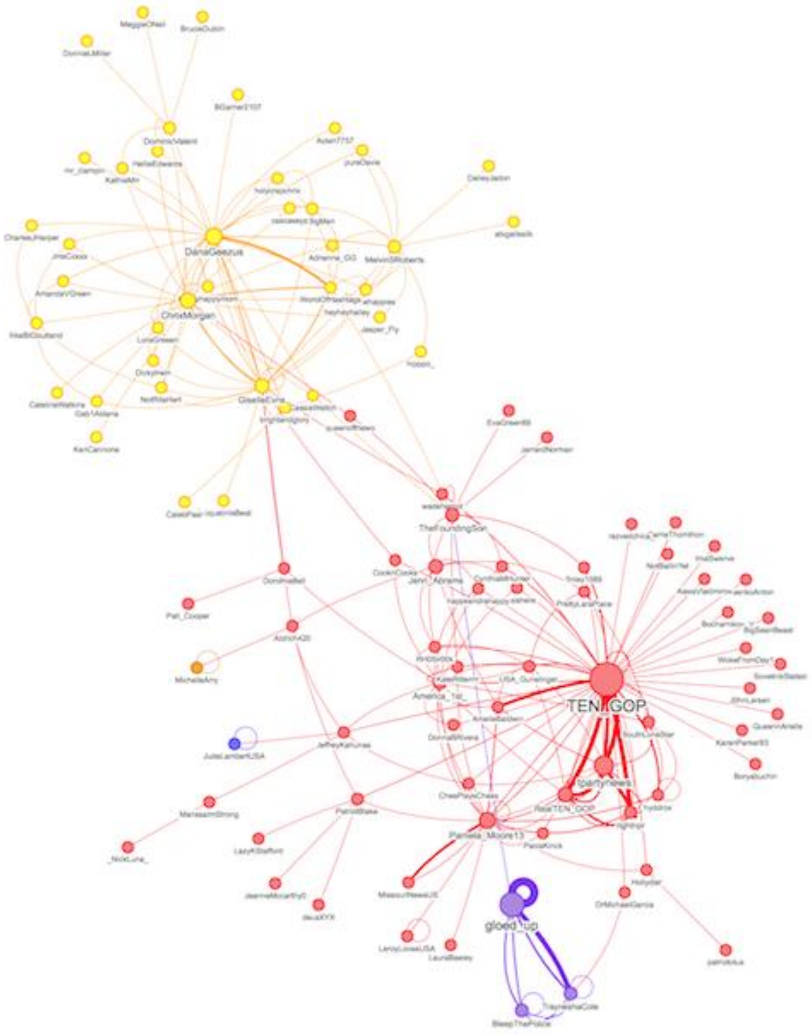


Figure 5. NBC used the community detection algorithm to show there are three communities in the Russian troll retweet network. Node size is proportional to the PageRank score for each node, which shows the importance of the account in the network.

Sarah Cohen is one of the pioneers who has been applying social network analysis in journalism since around twenty years ago. Rather than focusing on the visualization side of network analysis, she mainly used social network analysis as a reporting tool to decide angles and make breakthroughs in investigative stories.

In 2013, The New York Times did a story to examine a gun purchase website and questioned whether many gun sellers were essentially functioning as unlicensed firearms dealers, against federal law. They got advertisements from Armslist.com, a sprawling free classified ads Web site for guns. This website didn’t make a list of all the guns posted by a single seller, but each post included some links to other guns for sale by the same seller. Even if journalists couldn’t get a unique ID for the user and the sales history, they used the connected component algorithm, one of the community detection algorithms, to discover connections among different posts and found that a sprawling group of more than two dozen people had posted more than 20 different guns for sale in a several-month span.

“We scraped the website every night and used those ‘other posts from this user’ as a way to build what was called connected components, which was kind of like a daisy chain.” She said, “We used the connected component concept in SNA to tie them together. It’s more about using this concept to help find stories, rather than to display them.”

Challenges and Limits of Social Network Analysis in Journalism

In general, four limits of social network analysis, based on the research interviews, are as follows:

1. Data collection is one of the biggest challenges in conducting social network analysis. Sometimes the data collection itself would take up the majority of the time and render the following network analysis irrelevant.

First, it is hard to exhaust all the relevant sources to build a relational database. If the data is not stored digitally, journalists have to type it manually. Most of the time, scraping is also required at the first stage. Secondly, sometimes they have to extract relations from the unstructured data. The relation extraction algorithms are very unreliable when applied to the unstructured datasets. Thirdly, the graph database that comes from multiple data sources tend to be large and messy, even the path-finding and other connectivity-based algorithms sometimes produce unsatisfactory results, according to Stray in Network Analysis in Journalism: Practices and Possibilities.

2. What's more, a time-consuming, technical input of social network analysis sometimes cannot always guarantee the output. Most newsrooms rely on a well-established workflow that produces predictable content, for cost-efficiency is a major concern in newsrooms.

For example, using social network analysis does not guarantee a major breakthrough. It's possible to merely confirm some known connections or key players after reporters spend lots of time and effort running different algorithms.

"It's a reporting tool. Sometimes it's useful. Sometimes it's not. Generally, we have to collect data ourselves. We don't really need to use social network analysis, because all we need is a good way to keep notes. Once you collect the information, let's say 500 cases, it's pretty easy to know how people are connected--to see it without really having to do much else." Cohen said.

"A network graph isn't necessarily showing you something you don't already know." Peter Aldhous, science reporter from Buzzfeed said, "If you're having to compile a network of connections by hand, by the time you've done that reporting, you kind of already know what the network is going to show you."

"Analyzing relationships in your community can help you see how connections can translate into avenues of communication and possible sources of power for individuals within the network. But just like other methods of computer-assisted reporting, social network analysis is just a tool. It isn't going to tell you the whole story, but using it as a reporting tool can be a great place to start. By seeing relationships in a new way you might uncover possibilities within your investigation that you never imagined.” Dowdell said.

3. Most importantly, it may be difficult to humanize and interpret the algorithms result without sufficient contextual reporting. Especially in social network analysis, the outcome is usually in the form of a hairball of connections that may confuse the readers at first sight. Besides, the underlying meaning of its metrics, such as centrality and betweenness, is not easy to understand for those who have no background knowledge in this area.

Data journalists tend to use social network analysis as a reporting tool to serve the story itself. Graph algorithms can help uncover hidden clues and create network maps automatically, but can never take place of journalistic judgment and human intelligence. Social network analysis can never replace the role of journalistic judgment. Stories always come first. Graph algorithms come second. Journalists need to conduct more interviews to contextualize the story and weave a readable narrative.

“There’re metrics that you can apply. I kind of wouldn’t use it as a tool to get an answer from.” Aldhous said, “I’m sure there will be occasions when you’re doing social network analysis, you know that somebody is really important, but a pure network analysis possibly isn’t telling you that. Possibly what you do in that network doesn’t capture all the importance of that personal things.” He did a story to find the most influential players in cellular reprogramming by mapping out the citation network. But before he conducted the social network analysis and other statistical analysis, he had already known Shinya Yamanaka was one of the most important researchers in that field, based on his research and reporting, so he deliberately put Yamanaka in the center of the graph at first.

Connected China by Reuters is a noteworthy collaboration between journalists and social network designers. Dozens of journalists dug into government websites, policy papers, mainland China major publications, English news reporting, academic articles, and think-tank report to build their own database. Then they cooperated with a design firm based in Boston to produce the graphic.

Mark Schifferli, the project designer and data visualization expert from Fathom Information Design, worked closely with Irene Jay Liu and other Reuters journalists on this project. Schifferli said they needed journalists' judgment and familiarity with the Chinese civil service to help them determine the importance of people's political influence and affinity, "We were leaving it to journalists to characterize the data. That's more of a categorical choice they made for the specific relationships. We tracked everything that is the first-degree connections. The nature of that connection was specified by Reuter journalists."

But Liu's available hours to work in the Boston-based design team were limited. In order to scale the importance of officials' political influence, they configured the weights of different ranks by calculating the prominence of their careers and the strength of their ties to other important people. Then they used that importance score to size an icon for an official.

The closeness was often reported as "liked by", "reportedly close to", "mentor to" and so forth. They also used an algorithm that traverses the relationship in the network, looking at the links they have in common to indicate the possible closeness.

"The problem with social network analysis is how you distinguish between important connections and unimportant connections. You have hypothesis at first--what is important and what is not. My hypothesis is all of Xi's kindergarten connections are important. Without that, you can't just explore it. It's going to take you forever. It's more of a journalistic question." Chua recalled.

But they didn't track Xi's kindergarten connections in Connected China at that time. They painted a broad picture of the interpersonal relationships of Chinese elite politicians, instead of having more focused, journalistic assumptions and wrote stories based on some specific types of connections. Different from the Influencers project by ICIJ, the textual stories in Connected China are some individual profiles or analysis of a general political trend in China that can do without the database itself.

"If you want to tell a story, you should really look at the connection between Xi and this person you may not have known because they went to kindergarten or something. Then you have to make it clear for people, so they can follow it. That's very different from building an open-ended database you can explore." Chua said, "Probably the problem with a site like this is it's too powerful and let you do too many things and it didn't try to pull you down into very specific things that you may be looking for."

Therefore, we need to bring human intelligence and journalistic experience together to know what to connect. For example, when Chua was covering the Philippines, he focused on if a person was in the military, what year he graduated from the military academy, what fraternity he was in college and who was the godfather in his wedding. It would add more value to Connected China project to bring more China-specific, characteristic connections, based on the political and social norms in China.

4. Another challenge is that the social network database can be difficult to maintain. For example, initially, Chua planned to build a long-lasting, structured tool that could be updated for a long period of time. But Connected China was soon blocked by the Chinese government and got frozen since most of the team members left Reuters later.

Most of the social network projects were intended as archives for journalists and academic researchers to dig into, but the website would sometimes get frozen when project leaders jump to the next task or the funding is in shortage. A similar project is Poderapedia by Miguel Paz, which reveals links among Chilean business and political leaders.

“The idea of journalism is to make sense of the world and communicate it to the audience because people are busy, they live their lives, they don’t have time—you know, journalists spend time all day learning something and communicating it in 500 words. So there’s an editorial role. And I hate to say this, but yes, algorithms are very good at surfacing facts and putting them together. But that role of asking questions--computers are not good at that. They’re good at answering questions. And so the role of journalists will never change in that way, and the role of news organizations as the intuition and the infrastructure to enable individuals to be employed to do that work is not going to change.” Liu said in a journalism conference when asked to compare algorithms and editorial judgment.

Tips for Data Journalists

Below are several suggestions that could help overcome the shortcomings of social network analysis and maximize its advantage.

1. First, data journalists should have more collaboration in building a relational database and sharing open data. NBC News opened sourced the Twitter data in order to smooth the path for other journalists to further investigate the Russian influence in the 2016 election. Lots of local reporters took advantage of this dataset, according to Lyon at Neo4j. For instance, Buzzfeed published the TrumpWorld data that logged more than 1,500 people and organizations connected to the Trump administration. We should have more open source platforms that allow journalists and volunteers to add individuals and organizations of their knowledge and then suggest relationships.

2. Second, journalists should keep an eye on the development and application of social network analysis in other disciplines. Social network analysis combines the research methods of computer science, mathematics, statistics sociology, behavioral science and has been widely used in law enforcement and crime investigation and marketing. For example, marketing directors can use it to guide the promotion of their products and track feedback. Journalists, editors and newsroom managers can get inspiration from them to further make breakthroughs in news stories, maximize the influence of their newspapers and increase the readership.

3. Third, we should combine our solid contextual reporting with the social network analysis. Research and interviews may give us insight and angles which social network analysis fails to capture. “In journalism, in general, we don’t do very sophisticated analysis, we do fairly simple analysis. A lot of what we bring to it is the contextual reporting around the analysis. That’s why I think it (social network

analysis) is a little bit limited." Aldhous said, "It's a balance between what a formal network analysis tells you and what a wider contextual reporting tells you. In most cases in my experiences, you might do a little bit network analysis and it's the reporting around that that may give you the story."

4. What's more, when it comes to social network analysis tools, it's better to understand the underlying logic behind algorithms than keep up with different tools. Social network tools change very quickly. For example, very few journalists use UCINET, which was the mainstream analytic tool ten years ago. But the centrality and cluster algorithms behind those tools have long-lasting and sustainable application.

5. Journalists should establish a system and workflow for using social network analysis in newsrooms. Those reporters who're interested in this method can follow certain procedures and achieve more effective collaboration with teammates.

6. Last but not least, when reporters and editors wonder whether it's worthwhile to use social network analysis, here are several tricks that may help them make a decision: (1) Determine the nature of your story: if it's a story that mainly revolves around a large amount of nodes and links, or entities and relationships, social network analysis would be a good start to help you break ground. (2) Check the data availability: if it takes a much longer time to manually collect the relational data than the attribute data, you may need to think about whether the relational data is really essential in the story. (3) Conduct research and interviews with stakeholders: it can help you get a general picture about the network you're going to investigate, come up with some basic questions, and consider whether traditional journalistic methods would suffice to test your hypotheses. (4) Experiment with some social network analysis tools: some cutting-edge tools allow journalists to benefit from the graph

algorithms analysis in a user-friendly way. Both Neo4j and Gephi have embedded graph algorithms and simple interfaces. After importing your data into those tools, an interactive exploration of graph database or inspection of a network graph would give you some clues about what to connect and where to start.

APPENDIX

Appendix A: List of data journalists interviewed

Name	Organization	Job Title	Interview Method
Sarah Cohen	Arizona State University	Knight Chair of Data Journalism	Phone Interview
Brant Houston	University of Illinois	Knight Chair in Investigative and Enterprise Reporting	Skype Interview
Jaimi Dowdell	Reuters	Data Journalist	Phone Interview
Peter Aldhous	BuzzFeed	Data Journalist	Phone Interview
Reg Chua	Reuters	Chief Operating Officer	Phone Interview
Mark Schifferli	Fathom Information Design	Senior Developer	Skype Interview
Pierre Romera	ICIJ	Chief Technology Officer	Skype Interview
William Lyon	Neo4j	Leader of Neo4j Data Journalism Accelerator Program	Skype Interview
Miguel Paz	Poderopedia	Data Journalist	Phone Interview
Marc Smith	Social Media Research Foundation	Social Scientist	Phone Interview

Appendix B: Useful Tool and Resources

- **Neo4j:** a graph database that enables organizations to unlock the business value of connections, influences and relationships in data. ICIJ used Neo4j to analyze and visualize the Panama Papers and Paradise Papers.
- **Gephi:** an open-source and free tool that visualizes all kinds of graphs and networks, including Social Network Analysis, Exploratory Data Analysis, Link Analysis and so forth.
- **NodeXL gallery:** These are network graphs created with NodeXL, a template for graphing network data in Microsoft Office Excel.
- **Linkurious**: a graph visualization tool that provides social network analysis primarily through graph visualization. It is aimed at democratizing graph visualization techniques to help organizations extract concrete insights from graphs. Reporters can build their own graphics by selecting nodes from the underlying data.
- **LittleSis:** a grassroots watchdog network connecting the dots between the world's most powerful people and organizations.
- **Muckety:** Muckety publishes network maps and news stories to document paths of influence between government, business and nonprofit affiliations... The ICIJ analyzed the leaked HSBC files and uncovered the profiles of the main protagonists of this huge financial scandal and revealed the role of countries which took part in this fraud. They did it using stories and by building an archive of the profiles.
- **Theyrule.net:** a social network visualization that provides a glimpse of some of the relationships of the US ruling class. It takes as its focus the boards of some

of the most powerful U.S. companies, which share many of the same directors. Some individuals sit on 5, 6 or 7 of the top 1000 companies. It allows users to browse through these interlocking directories and run searches on the boards and companies.

- **Opencorporates:** the largest open database of companies and company data in the world, with in excess of 100 million companies in a similarly large number of jurisdictions. It's a helpful database for journalists to tackle the use of companies for criminal or anti-social purposes, for example corruption, money laundering and organized crime.
- **Connected Actions:** Connected Action consulting group applies social science methods in general and social network analysis to enterprise and internet social media usage by collecting and analyzing data from Twitter, Facebook, message boards, blogs, wikis, friend networks, and shared file systems.
- **Graphika:** Graphika offers detailed and contextual social media analysis, surfacing audience insights and key content being shared and provides tools for understanding and participating in social media communities.
- **Influence Mapping Group:** Influence mapping documents and maps relationships between people, organizations, and political processes and share technologies that help structure, visualize and analyze influence networks.
- **Kumu:** a data visualization platform that organize complex information into interactive relationship maps, such as stakeholder mapping, systems mapping, social network mapping, community asset mapping.
- **IRE Tipsheets:** guides, tips, and presentations that involve social network analysis on the IRE website.

- **Peter Aldhous's tutorials:** Peter Aldhous's tutorials on social network analysis and Gephi.
- **Jonathan Stray's tutorials:** Jonathan Stray's network analysis tutorials that he taught at Columbia Journalism School.

Appendix C: Project Proposal

Introduction

Before I came to the Missouri Journalism School, I studied journalism at Wuhan University. In college, I attended the Dy Club Data Media Lab, participating in online chats given by prestigious data journalists. I also helped translate data journalism projects from ProPublica, FiveThirtyEight, and other news outlets. During several internships at domestic and foreign media outlets, I came to know the importance of data and graphics in the journalism industry.

At the Missouri Journalism School, I worked at the IRE database library and enrolled in various courses to further enhance my understanding about data journalism and strengthen my skills to work as a data reporter. The Computer-Assisted Reporting class taught by Professor David Herzog gave me an opportunity to submit a Sunshine request and negotiate with government departments for public records. I also learned to use SQL (Structured Query Language) to manage and analyze different datasets. In the Investigative Reporting class taught by Professor Mark Horvit, I attempted to combine the data analysis and visualization with narratives and storytelling. In the Advanced Data Journalism class taught by Chase Davis, I learned to use Python to scrape structured data from a website and utilize Python libraries, such as Matplotlib, to visualize large-volume datasets. In the Multimedia Planning and Design class taught by Rob Weir, I learned HTML, CSS and JavaScript to design websites and create interactive graphics. At the IRE database library, I filed Freedom of Information Act requests, used My SQL and R to update databases, checked the integrity of datasets, analyzed databases, and created interactive maps and graphics for clients.

I have developed the knowledge and skill of data visualization and analysis from various courses and jobs, which primes me for conducting this graduate project. I'm interested in studying the role of network graphs in analyzing network relationships. I will take advantage of what I have learned from the Missouri Journalism School and apply that knowledge to the summer internship at USA Today while working on my graduate project.

Professional skills component

This summer I will intern at USA Today from June 4 to September 7. During my 14-week internship, I will spend 32 hours per week contributing to data visualization and analysis, under the supervision of the interactive graphics editor Shawn Sullivan. I explained my skills and passion about working on a long-term, investigative project and Shawn Sullivan said I might have the opportunity to follow a long-term project of my interest during the summer.

At the same time, I will work on my graduate project. I will contact and conduct interviews with the data reporters and editors who have been working on network visualization projects, especially "Connected China"(Liu et al., 2013) by Reuters and "The Influencers" (Romera et al., 2017) by ICIJ, getting to know the techniques and details about visualizing power relationships and provide advice to other journalists who are also interested in this topic.

Theoretical framework

Social network theory

Social network analysis is a method of research that uses data to analyze the relationships between actors in a network. It is grounded in systematic relational data and draws heavily on graphic imagery and computational models. The node refers to the actors or subjects, such as individuals, organizations or nations. The edge, also known as the tie or link, describes the relationship, such as family ties, business cooperation and political alliances.

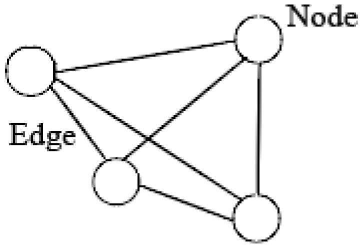


Figure 1. Nodes and edges in social network analysis

There are mainly two types of data: attribute and relational (Scott, 2000). Attribute data refers to the properties and characteristics of an individual, while relational data refers to the ties between people. Aimed at obtaining higher-level descriptions of the structure from the low-level relational data, the network analysis focuses on the relationships between individuals and social structures, instead of the specific characteristics of people (Rice & Richards, 1985).

Social network analysis is a means of investigating social structures. It conceptualizes social structure as a network with connected members and exchangeable resources (Otte & Rousseau, 2002). Social network analysis is also a set of approaches or techniques to study the exchange of resources among various actors such as the individuals, groups, or organizations (Haythornthwaite, 1996). It helps us determine what and how resources flow from one actor to another. Communication networks, according to Shumate and Contractor (2013), are relationships between various types of actors that illustrate the ways in which messages are transmitted, exchanged, or interpreted.

Another important aspect is to study how the structural relationships affect the behaviors and beliefs of actors within the network. The structure of relationships among actors and the locations of actors in the network have important behavioral, perceptual and attitudinal consequences for the units and for the system (Knoke & Kublinski, 1982).

The two types of social networks include the whole network, which offers a view of the entire structure of the environment, and the egocentric network, which offers a view of the network from the perspective of an actor within the network (Scott, 2000).

The interest in human networks began to emerge in the research of anthropologists, socio-metricians and sociologists. (Berkowitz, 1982; Coleman 1958; Rogers & Kincaid, 1981). The majority of their study focuses on the community power structures, organizational communication and corporate directorates and the like.

The recent development of computer science and graphical techniques contributed to enlarging the scope and enriching the methods of network analysis. Through the application of graph theory, some basic concepts, such as the "distance" between two individuals, the relative "centrality", the formation of "cliques" and the "densities" of the whole network, can be detected and quantified (Scott, 1996).

Scott (1992) also defined the following steps to collect, sample and organize the relational data.

(1) Collection

The relational data can mainly be collected from documentary sources, ethnographic investigations and surveys. The ethnographic investigations use observation and conversation to collect data.

(2) Sampling

The boundary of the social network needs to be defined. For example, the identification of top leadership, justification for the cut-off threshold to define the “top-level”. As researchers may have inaccurate views about the boundaries of relational systems, it is possible that the social network studied will be an imperfect representation of the full network, especially in the case of informal groups.

(3) Organizing

Two models can be used to represent the data: similarity model and linkage model (Rice & Richards, 1985). The similarity model involves position or distance, using the presence or absence of a link between two nodes as data, represented by 1 or 0 in the corresponding matrix cell. The linkage matrix uses scalar or binary values to show relationships or cohesion, of which the higher values refer to highly-linked nodes.

Literature review

Social Network Visualization

The social network visualization can not only help illustrate the complicated relationships in a large social network, but also provide in-depth insights about the power structures and organizational communication. There are two main ways to represent the network graphs: a graph made up of multiple nodes and edges or a matrix where rows and columns represent individuals and the numbers in each cell stand for their relationships (Viégas & Donath, 2004). The first way is more vivid, legible and is widely applied to display the network relationship.

The application of graphics in social network analysis started from the 1930s and gradually flourished with the spread of personal computers and World Wide Web (Freeman, 2000). One of the earliest known social network graphics is the Friendship Choices Among Fourth Graders by Jacob Moreno in 1934, in which he used triangle nodes to represent boys and the circle nodes to represent girls.

Before the emergence of computer-generated graphics, some scholars developed the directed graph (a set of nodes connected by edges, where every edge has a direction) and target sociogram (centrality is maximum at the center and decreases with distance from the center), specified the nuclei of a network, used computational procedures to place points, and used multidimensional scaling to locate points and map three-dimensional arrays (Scott, 1992). After the 1970s, with the development of computational languages, Alba (1972), Lesniak, Yates, Goldhaber and Richards (1978) developed the SOCK, NEGOPY, ORTEP and View_Net programs to produce network images automatically.

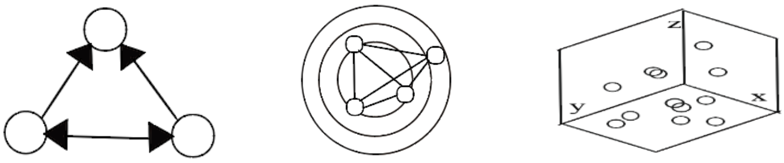


Figure 2. Directed graphs, target sociogram and three-dimensional graphs.

In the network, actors are represented by the points and the relationships are represented by the lines. These two elements also define a graph (Harary, 1959). Graph theory is used to describe the patterns of connections among points.

Some software tools have developed the accessible digital interfaces for users to visualize social networks, such as Gephi, Neo4j, GUESS, PAJEK, UNICET.

Social Network Analysis in journalism

Social network analysis has increasingly been used in organizational communication and computer-mediated communication (Monge & Contractor, 2003; Shumate et al.,2013), and contributes to theoretical innovations in journalism, such as the emergence of the third-level agenda setting (e.g., Guo, Vu & McCombs, 2012). Fu (2016) found the current research mainly concentrated on the semantic and flow networks and its application to journalism studies has yet to be fully realized.

Social networks are widely used to identify clusters of public attitudes. Katz and Lazarsfeld (1955) found messages from the media may be further mediated by opinion leaders who interpret and diffuse the information to the social networks in which they are embedded. The emerging social media platforms, such as Facebook, Twitter, LinkedIn, offer abundant social network resources for journalists to find stories and collect information about public opinion. Based on the social network analysis, journalists are able to visualize and analyze the patterns of interaction. John Kelly (2008) used a social network diagram to map the English language blogosphere and discover the major clusters around politics and technology. In the networked public sphere, clusters form around the shared concerns and similar sources, and a network of social knowledge is knit together.

Social network analysis is also a useful tool for analyzing the power structure and relationships (Kerbo & Fave, 1979). The goal of social network analysis is to observe linkages through construction of visual patterns (Hansen, Shneiderman & Smith, 2011). A number of journalistic pieces use the social network data to visualize and analyze government-business collusion, informal factional networks and family ties. Some examples include Mayor Bloomberg’s Circles of Power, Oscar Contenders and Wen Family Empire by The New York Times, Zhou’s Power Base by Caixin, The Calderon family’s connection by The Los Angeles Times, and Spheres of Influence by The Washington Post. The Knight Lab's Untangled initiative aims to catalogue useful tools and investigations that use SNA , and unite communities interested in the application of SNA to journalism.

The centrality measures from social network analysis can help journalists find influential actors in a large network. Dowdell (2004) used SNA to map the network of powerful attorneys in Columbia, Missouri and found SNA a good way to visualize connections, analyze relationships and discover potential sources of power, especially in the area of local government power, campaign finance, crime and public health.

In order to answer the question “who’s important here,” below are some key concepts generally applied in the network visualization (Schutt, 2013): (1) degree: counts how many people are connected to you. (2) closeness: if you are close to everyone, you have a high closeness score. (3) betweenness: people who connect people who are otherwise separate. If information goes through you, you have a high betweenness score. There is a formula to calculate the score using the shortest distance between two actors. (4) eigenvector centrality: a high eigenvector score means that a node is connected to many nodes who themselves have high scores. For example, a person who is popular with the popular kids has high eigenvector centrality. Some network software is very useful to compute the centrality measures: NetworkX or igraph for Python, statnet for R, and NodeXL for Excel.

This approach of leveraging computed attributes is particularly valuable for social network analysts to discover the pattern, distribution and trend, as the inherent attributes, such as the gender and age, do not tell the whole story (Perer, 2010). Chris Wilson, editor of US News & World Report, used SocialAction to conduct social network analysis and visualize voting patterns in the Senate. In 2007, he used betweenness centrality and clustering algorithms to uncover the gravity centers and geographic alliances among voters.

Methodology

This project will dig into the role and methods of network visualization to illustrate the power structure and relationship, through the case study of two successful network visualization projects --"Connected China" by Reuters, and "The Influencers" by ICIJ. The targeted publications include *Flowing Data, Poynter, The Pudding, IRE Journal and the Columbia Journalism Review.*

Based on the literature review, the research questions in this study include:

Research Question 1: What's the process for journalists to produce a social network visualization?

Research Question 2: What challenges and limitations do journalists encounter when using a social network visualization to analyze power relationships?

Case Study

Researcher Robert K. Yin (1984) defined the case study as an empirical inquiry that investigates a contemporary phenomenon within its real-life context.

I will look into two successful projects of network visualization: "Connected China" by Reuters in 2013, and "The Influencers" by ICIJ in 2017. These two cases both use social network visualization to illustrate the power dynamic, but they deliver different levels of message, and adopt distinctive visual narratives. As mentioned above, there are two types of social networks: the socio-centric network, which focuses on the whole group, and the ego-centric network, which consists of a focal node “ego” and the nodes to which ego is directly connected to.

From the macro-level perspective, “Connected China” sheds light on the whole political structures and faction ties. From the micro-level perspective, “The Influencers” represents the part of the network surrounding a single person. It is more microscopic, focusing on the inner circle of Donald Trump and emphasizes the image and resources of each individual.

“Connected China” is a detailed and comprehensive data visualization produced by Reuters to help decipher the social and institutional power of Chinese elite politics, including the social networks of its family, mentorship, rivalry, alliance. A team of dozens of journalists and researchers took one year and a half to visualize tens of thousands of entities, more than 30,000 relationships and 1.5 million words.

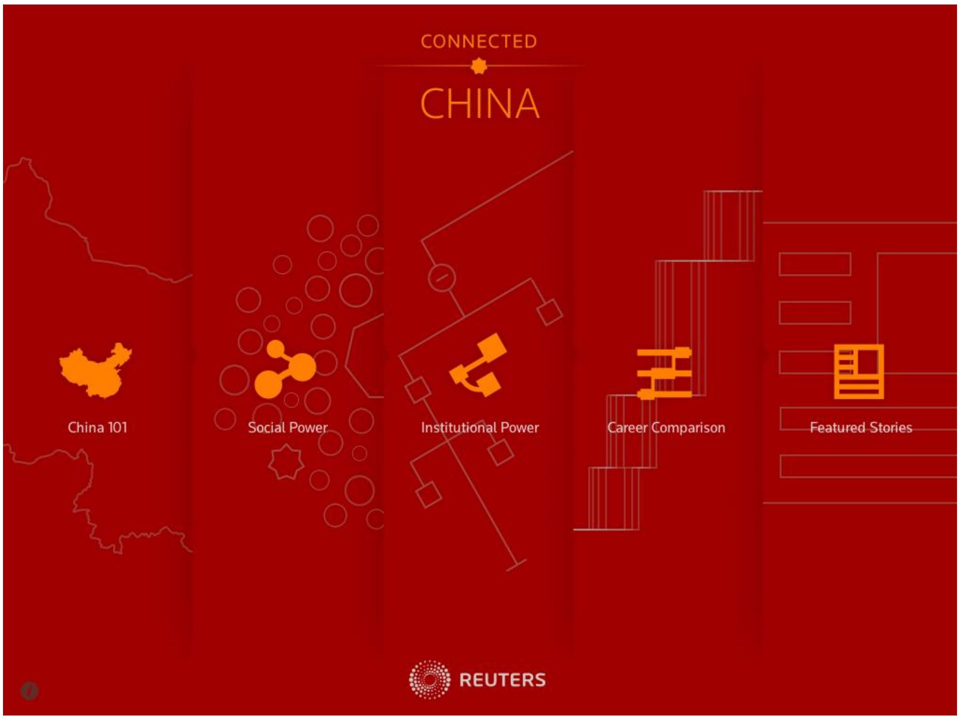


Figure 3. A screenshot from “Connected China”, a project produced by Reuters to explore the people, institutions and relationships that form China's elite power structure

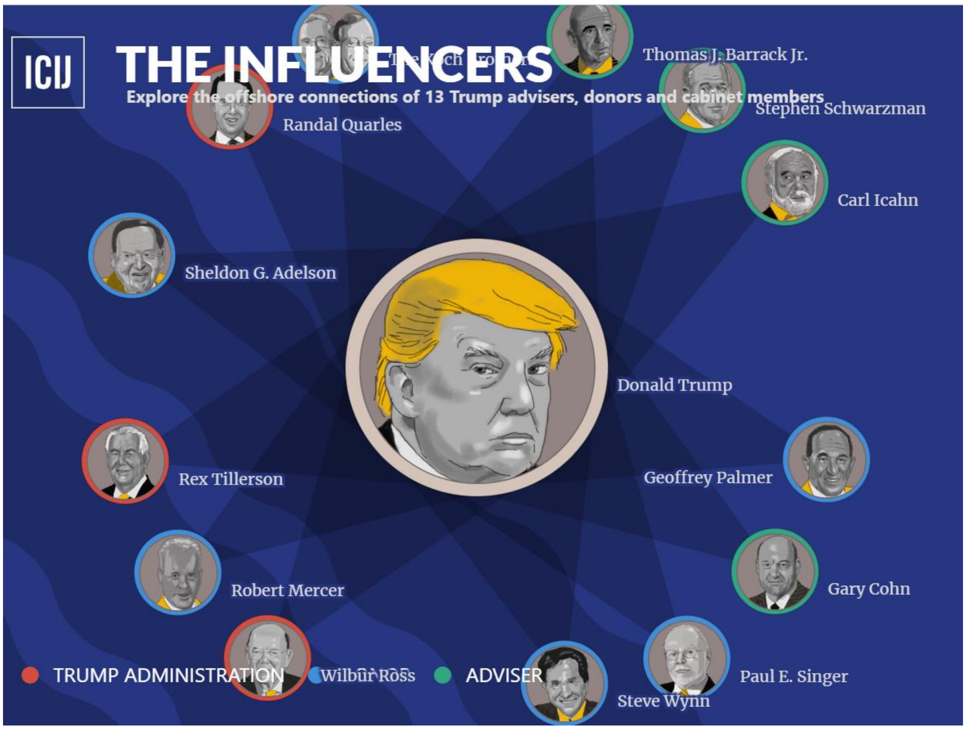
There are five sections of Connected China: China 101, Social Power, Institutional Power, Career Comparison and Featured Stories.

The first key section is Social Power: (1). Families: Princelings; Golden Sons-in Laws. (2). Coalitions: Shanghai Clique; Tuanpai. (3). Relationship with party leaders. (4). Accumulation of guanxi, which is the accumulated social capital.

The second key section is Institutional Power: (1). The Communist Party: the collective leadership by the Politburo Standing Committee;(2). The three pillars of Chinese politics: party, state and military

The third key section is Career Comparison: (1). the path to political power; (2). comparing the ages of political leaders; (3). the impact of retirement ages; (4). comparison among multiple individuals.

The second case I will study is "The Influencers", the network visualization of U.S. President Donald Trump's offshore connections from the "Paradise Papers" project. President Donald Trump is connected with a number of advisors, donors and cabinet members who have taken advantage of the offshore tax havens for their own business profit. "Paradise Papers" is a global investigation based on 13.4 million leaked files from leading offshore firms in 19 secrecy jurisdictions. ICIJ cooperated with 96 news agencies worldwide to reveal the offshore activities of some of the world's most powerful entities, including the financial transactions of 120, 000 people and companies.



THE INFLUENCERS

Explore the offshore connections of 13 Trump advisers, donors and cabinet members

Figure 4. A screenshot from "The Influencers", a project produced by ICIJ to explore the offshore connections of 13 Trump advisors, donors and cabinet members

Interview

According to Newton (2010), interviews have the following advantages for scholarly research:

(1) They provide the opportunity to generate rich data;

(2) The language used by participants is considered essential in gaining insight into their perceptions and values;

(3) Contextual and relational aspects are seen as significant to understanding others' perceptions;

(4) Data generated can be analyzed in different ways.

The structured interview is also known as a standardized interview or a researcher-administered survey, in which the same questions will be given to the interviewees in the same order to reduce context effects (Brinkmann, 2014). Context effect describes the influence of environmental factors on one's perception of a stimulus. The validity and reliability of the research data can be increased by the structured interview within the same context (Sudman & Gullickson, 1997). On the other hand, the structured interview lacks the interactive nature of communication (Opdenakker, 2006).

The unstructured interview has no prepared questions and the interviewer has relatively less control in the conversation (Ryan et al., 2009). However, the unstructured interview contributes to building trust and creating a positive environment for the interviewee to share opinion (DiCicco-Bloom & Crabtree, 2006).

I chose the semi-structured interviews to absorb the advantages of both the structured interviews and unstructured interviews. The semi-structured interview means that researchers have a set of questions about a certain topic, but the conversation is free to vary and is likely to change substantially between participants (Fylan, 2005).

The interview in my study will include both closed-ended questions and open-ended questions. The answers of close-ended questions are always fixed in advance (Brinkmann, 2014). Open-ended questions require respondents to formulate a response in their own words (Zull, 2016), which allow the interviewees to have a more freewheeling discussion about the future directions of network visualization.

The semi-structured interviews will be conducted to explore the workflow and challenges of the Connected China project and the Influencers project. I will interview the main contributors of Reuters and ICIJ and some network visualization journalists for their opinion and advice on using social network analysis in the newsroom.

I will contact the interviewees via email in advance, scheduling the time and location, and conduct the interviews in person or via Skype if that is not possible. Each interview will last between 30 to 60 minutes. Since this project is carried out over a period of three months, I have some chances to adjust the questions and strategy based on the completed interviews.

Then I will transcribe the interviews and extract the key information. I will summarize the workflow to produce social network visualizations, evaluate the pros and cons and provide advice to other journalists who are interested in using social network analysis.

Interview Guide

The interview questions are framed and developed according to the research question. The potential interviewees are chosen from the main contributors of "Connected China" and "The Influencers" and other established data visualization journalists and developers.

Research Question 1: What’s the process to produce a network visualization?

- What’s your general workflow?
- How did you collect, clean and store the relational data?
- What metrics or computed attributes of social networks did you use? What journalistic questions did you try to answer?
- What software or programming languages do you recommend?
- What kinds of narrative devices, visual structure and interactive elements did you consider when visualizing the relational data?

Research Question 2: What challenges and limitations do journalists encounter when using the social network visualization to analyze power relationships?

- What are the difficulties you’ve faced in the process? How did you solve them?
- What are the limitations of the current project? What are the potential updates in the near future?
- What are the alternate methods for journalists to investigate social networks, besides using social network analysis?
- What are the effect and feedback of your project? How did it affect your reporting?
- What lessons did you learn from this project? What advice would you give to other journalists and designers on visualizing social networks?

Below are some potential interviewees, who have contributed to the aforementioned projects, or had extensive experience in the area of social network visualization:

- Reg Chua, the chief editor of the Connected China project and data and innovation editor of Reuters
- Irene Jay Liu, the project director of Connected China and news and data editor of Reuters
- Ben Fry, the project designer and data visualization expert from Fathom Information Design
- Other members of Connected China: Production heads Yolanda Ma and Malik Yusuf; Copy editor John Newland etc.
- Victor Shih, associate professor of political economy and Chinese elite politics at US San Diego University and co-founder of Communist Party Elite Database.
- Sasha Chavkin, Spencer Woodman, Martha M. Hamilton, Emilia Díaz-Struck, reporters and editors of ICIJ
- Pierre Romera, lead developer of ICIJ; Rocco Fazzar and Javier Arce (Populate Tools), designers of ICIJ
- Michael Hunger & William Lyon, developers at Neo4j who have been working on several network graph projects, such as Panama Papers, Paradise Papers.

- Sarah Cohen, Knight Chair in Data Journalism, former data reporter and editor of the New York Times and the Washington Post, who has used SNA in various projects

- Brant Houston, Kight Chair in Investigative and Enterprise Reporting, former director of IRE, whose CAR textbook included an appendix on SNA

References

Baumgartner, T., Buckley, W., Burns, T. R., & Schuster, P. (1976). Meta-power and the Structuring of Social Hierarchies.

Berkowitz, S. D. (2013). *An introduction to structural analysis: The network approach to social research*. Elsevier.

Brinkmann, S. (2014). Interview. In *Encyclopedia of critical psychology* (pp. 1008-1010). Springer New York.

Bounegru, L., Venturini, T., Gray, J., and Jacomy, M. (2016). Narrating Networks: Exploring the Affordances of Networks as Storytelling Devices in Journalism. *Digital Journalism*.

Cartwright, D., & Harary, F. (1956). Structural balance: a generalization of Heider's theory. *Psychological review*, *63*(5), 277.

Coleman, J. (1958). Relational analysis: the study of social organizations with survey methods. Human organization, 17(4), 28-36.

DiCicco-Bloom, B., & Crabtree, B. F. (2006). The qualitative research interview. *Medical education, 40*(4), 314-321.

Dowdell, J. (2004). *Journalism and Social Network Analysis: Visualizing Power and Influence*. (Unpublished master's thesis). University of Missouri-Columbia, Columbia, Missouri

Etling, B., Kelly, J., Faris, R., & Palfrey, J. (2010). Mapping the Arabic blogosphere: Politics and dissent online. *New Media & Society*, *12*(8), 1225-1243.

Freeman, L. C. (2000). Visualizing social networks. Journal of social structure, 1(1), 4.

Fu, J. S. (2016). Leveraging Social Network Analysis for Research on Journalism in the Information Age. *Journal of Communication*, *66*(2), 299-313

Fylan, F. (2005). Semi-structured interviewing. *A handbook of research methods for clinical and health psychology*, 65-78.

Guo, L., Vu, H. T., & McCombs, M. (2012). An expanded perspective on agenda-setting effects: Exploring the third level of agenda setting. *Revista de Comunicación*, (11), 51-68.

Harary, F. (1959) Status and contrastatus, Sociometry 22, 23-43

Harary, F., Norman, R. Z., & Cartwright, D. (1965). Structural models.

Haythornthwaite, C. (1996). Social network analysis: An approach and technique for the study of information exchange. Library & information science research, 18(4), 323-342.

Himelboim, I., Smith, M. A., Rainie, L., Shneiderman, B., & Espina, C. (2017). Classifying twitter topic-networks using social network analysis. *Social Media+ Society*, *3*(1), 2056305117691545.

Liu I. J., Chua R., Fry B., Ma Y., Yusuf M., Newland J., … Schifferli M. (2013, February 28). Connected China. *Reuters*, Retrieved from http://china.fathom.info/

John W.P. & Elizabeth M.P. (2017). *Network Theory and Political Science*. Oxford: Oxford Handbooks.

Katz, E., & Lazarsfeld, P. F. (1955). Personal influence: The part played by people in the flow of communications. *Glencoe, IL: Free Press of Glencoe*.

Kelly, J. (2008). Pride of place: Mainstream media and the networked public sphere.

Kelly, J. (2008). Mapping the blogosphere: Offering a guide to journalism's future. *Nieman Reports*, *62*(4), 37-39.

Kerbo, H. R., & Fave, L. R. D. (1979). The empirical side of the power elite debate: an assessment and critique of recent research. *The Sociological Quarterly*, *20*(1), 5-22.

Knoke, D. & Kublinski, J.H. (1982). *Network Analysis*. Beverley Hills: Sage Publication.

Krebs, V. E. (2002). Mapping networks of terrorist cells. *Connections*, *24*(3), 43-52.

Lindlof, T. R., & Taylor, B. C. (2017). *Qualitative communication research methods*. Sage publications.

Luce, R. D. (1950). Connectivity and generalized cliques in sociometric group structure. *Psychometrika*, 15(2), 169-190.

Luce, R. D., & Perry, A. D. (1949). A method of matrix analysis of group structure. Psychometrika, 14(2), 95-116.

Newton, N. (2010). The use of semi-structured interviews in qualitative research: strengths and weaknesses. *Exploring qualitative methods*, *1*(1), 1-11.

O'Neil, C., & Schutt, R. (2013). *Doing data science: Straight talk from the frontline*. " O'Reilly Media, Inc.".

Opdenakker, R. (2006, September). Advantages and disadvantages of four interview techniques in qualitative research. In *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research* (Vol. 7, No. 4).

Perer, A. (2010). Finding beautiful insights in the chaos of social network visualizations. *Beautiful Visualization: Looking at Data Through the Eyes of Experts*.

Rice, R.E., & Richards, W.D. (1985). An overview of network analysis methods and programs.

Rogers, E.M. & Kincaid, D.L. (1981). Communication Networks: Toward a New Paradigm for Research. New York: Free Press.

Romera P., Chavkin S., Woodman S., Hamilton M. M., Fazzari R., Arce J., … Hillhouse J. (2017, November 5).The Influencers. *International Consortium of Investigative Journalists*, Retrieved from https://projects.icij.org/paradise-papers/the-influencers/#/

Scott, J. 1992. Social Network Analysis: A Handbook. London: Sage.

Scott, J. (1996). Software review: A toolkit for social network analysis. Acta sociologica, 39(2), 211-216.

Seidman, S. B., & Foster, B. L. (1978). A graph-theoretic generalization of the clique concept. *Journal of Mathematical sociology*, *6*(1), 139-154.

Shumate, M., & Contractor, N. (2013). Emergence of multidimensional social networks. *The SAGE handbook of organizational communication*, 449-474.

Smith, M. A., Rainie, L., Shneiderman, B., & Himelboim, I. (2014). Mapping Twitter topic networks: From polarized crowds to community clusters. *Pew Research Center*, *20*, 1-56.

Stray, J. (2017, August). *Network Analysis in Journalism: Practices and Possibilities,* presented at the 23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), Halifax, Canada.

Stray, J. (2011, August 1). *Visualizing Communities*. Retrieved from http://jonathanstray.com/visualizing-communities

Sudman, S., Bradburn, N. M., Schwarz, N., & Gullickson, T. (1997). Thinking about answers: The application of cognitive processes to survey methodology. *Psyccritiques*, 42(7), 652.

Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2).

Venturini, T., Bounegru, L., Jacomy, M., & Gray, J. (2015). How to Tell Stories with Networks: Exploring the Narrative Affordances of Graphs with the Iliad.

Venturini, T., Jacomy, M., Bounegru, L., & Gray, J. (2017). Visual Network Exploration for Data Journalists.

Venturini, T., Jacomy, M., & Pereira, D. (2015). Visual Network Analysis: The Example of the Rio.

Viégas, F. B., & Donath, J. (2004, November). Social network visualization: Can we go beyond the graph. In *Workshop on social networks, CSCW* (Vol. 4, pp. 6-10).

Otte, E., & Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. Journal of information Science, 28(6), 441-453.

Ward, M. D., Stovel, K., & Sacks, A. (2011). Network analysis and political science. *Annual Review of Political Science*, *14*, 245-264.

Wasserman, S. & Faust, K. 1994. Social Network Analysis: Methods and Applications. Cambridge: Cambridge University Press.