

SOCIAL NETWORK ANALYSIS IN JOURNALISM: VISUALIZING POWER RELATIONSHIPS

Chen Chang

David Herzog, Project Supervisor

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 Asiatici, 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.

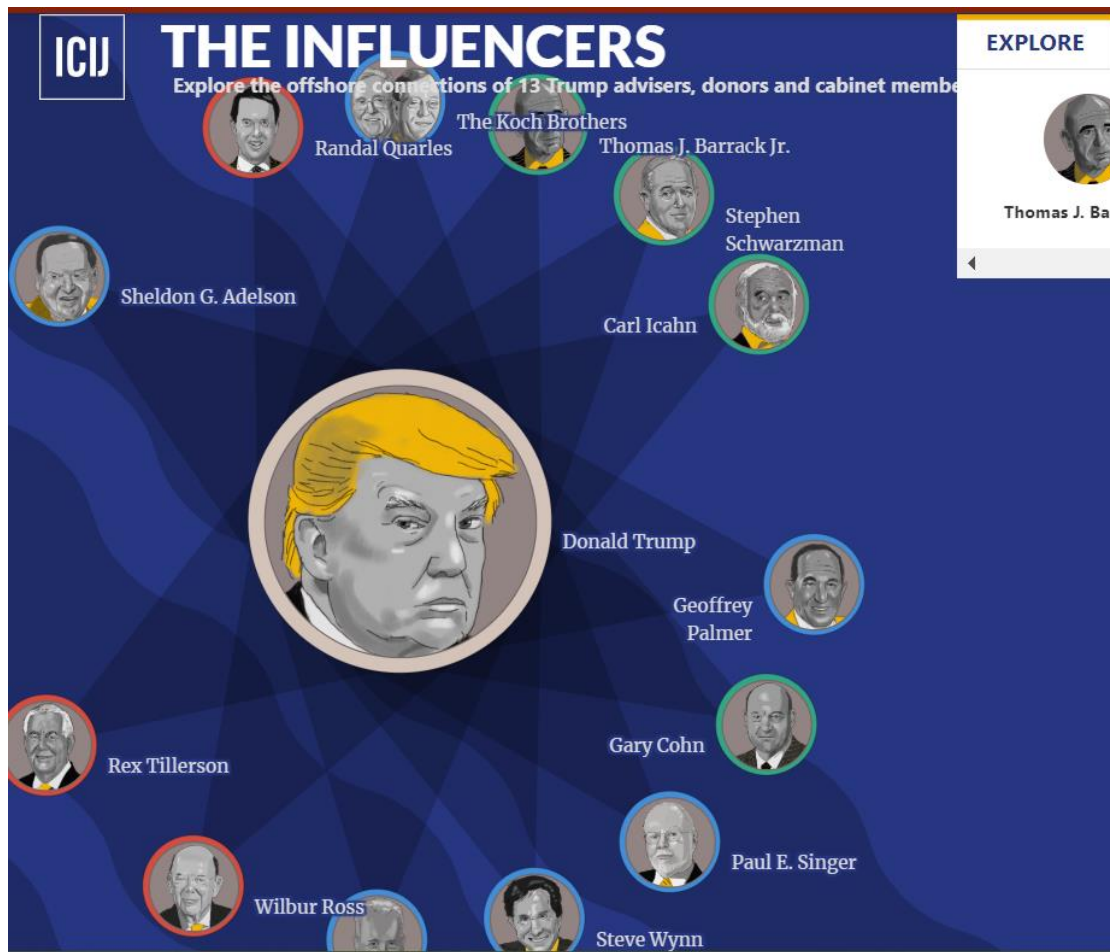


Figure 1. A screenshot from “The Influencers”, showing Trump’s social connections.

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.

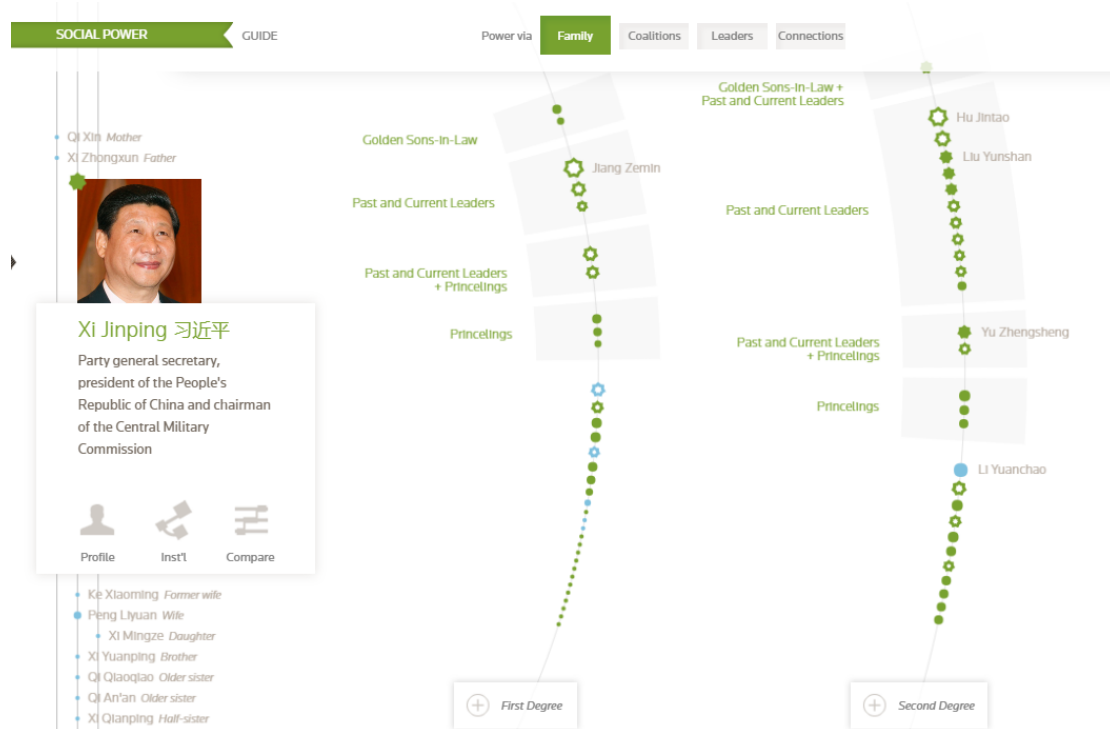


Figure 2. A screenshot from “Connected China”, showing Xi Jinping’s social connections.

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.

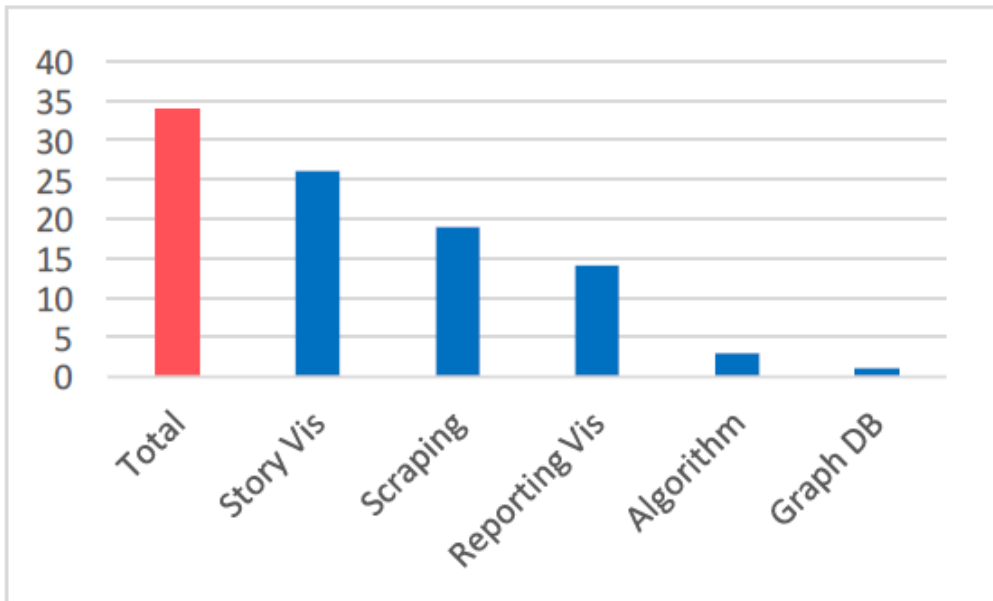


Figure 3. Attributes of the 34 network analysis stories collected by Jonathan Stray in *Network Analysis in Journalism: Practices and Possibilities*.

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

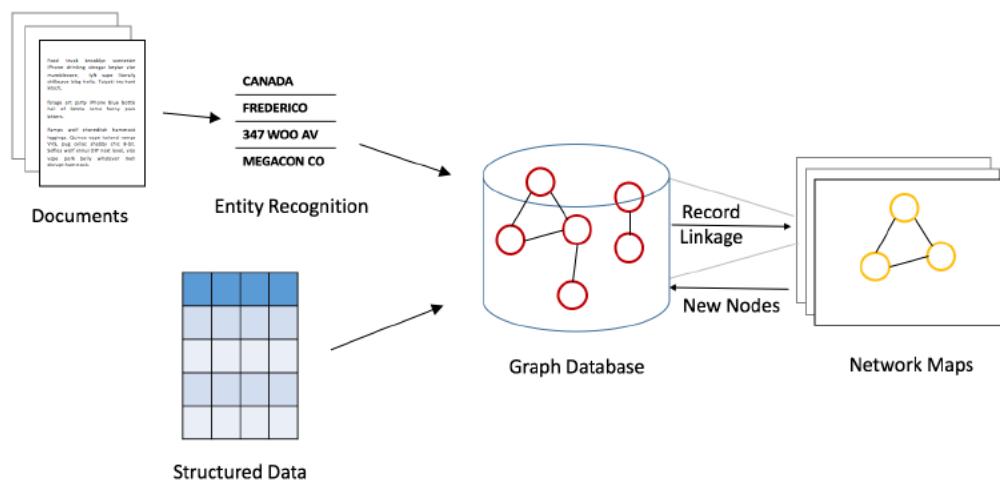


Figure 4. The proposed design for an integrated system for network analysis in investigative journalism by Jonathan Stray.

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.

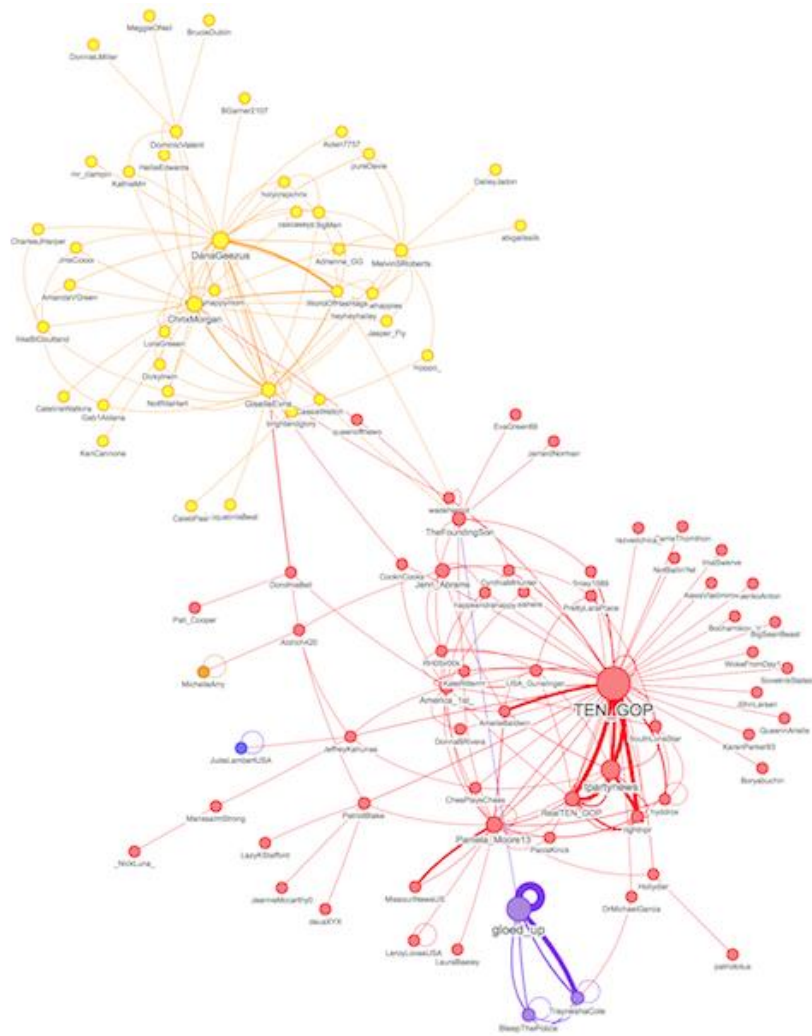


Figure 5. NBC used the community detection algorithm to show there are three communities in the Russian troll retweet network. Node size is proportional to the PageRank score for each node, which shows the importance of the account in the network.

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.