These transcriptions may contain errors, especially in spelling of names. These are unfortunate, and we regret that we do not have the resources to fix these errors. Still we believe these transcripts will be valuable to many users.

Words and Networks: Considering the Content of Text Data for Network Analysis

Jana Diesner, UIUC

>> Jana Diesner: I am very happy that you all are here after your lunch break and in your very last session. I'll try to make this fun and entertaining for you. How many of you have ever done some sort of network analysis? Almost all of you. How many of you have ever done some sort of texted analysis? And how many of you have done both on the same project? Ah ha! How did it go? [Laughing] Did you hand code things? Did people who hand coded things get annoyed ever when doing that? [Laughing] That was me, that was me, that was the whole motivation. I was sitting there and I thought how cool would it be to not only do content analysis, where you count things in text data and see how often they occur, and then you make this weight assumption that frequency of occurrence translates into importance, debatable, but that's what content analysis is kind of about. So you do that, but you could also look at the connections between the things. It's not only that Clinton and Bush are mentioned but what's the relationship between them and what weight are they co-mentioned? And then try to [inaudible] by hand. Has anybody ever done that? Look for? Did you like it? Really? How many of you have seen the John Nash movie? The "A Beautiful Mind"? Who remembers what John Nash was doing in the barn, or in his little back house? What does he do? He did network text analysis. And what happened to John Nash? [Laughing] Do you remember? That was me too. At that point I said this is not going to happen by hand. This has to be automated in some way. And I'll talk about how we can take texted and make it useful for network analysis. All right. So here are two quotes. They are both from communications and scholars actually. One of them says, "We cannot reduce communication to as its transmission." What we often do-how many of you have built communication data networks? Like networks between email senders or something, or people in a meeting, or people who talk to each other on the phone. So what we often do there is we take the people who send-Mark sent an email to Ben, so we have Ben as a node, we have Mark as a node, and a link from Mark to Ben. This is how those networks often look like. And then we have all these communication partners and we can look at the network of communication interaction and do our classic network analysis. Well, what we often don't look at is what do they talk about? And then our other question is does that matter? Does it matter that we look into the content of communication or does it not? And honestly we don't have a lot of good answers to that. It sounds like a trivial question, PhD students pay attention, it's not a very good question. But we don't have a whole lot of hard answers to that. When does it actually make sense to consider the content of communication for looking at networks? I don't know. I have a few answers for my [indecipherable] work, but not a theory or something. And the other one says, "I'm traveling through the networks of [indecipherable]." Which could be, for example, language, knowledge, ideas, information. So whenever we have that, sometimes, we might want to not only consider the structural of who's talking to who, but also looking in what we are [indecipherable] and what are you talking about. And all my research fits kind of this model. So I do three things and I really do only work at the intersection of any two of them not all three of them. One of them is social science and network analysis. And what I mainly get from that domain, or what I use from that domain, are models and theories that give me some theoretical grounding of my work, and that I can use or I test some theories from that domain. But this is where my theory making comes from. Then I use natural language processing to analyze text data. And within the field of natural language processing, which is really large scale data mining, it can also be small scale actually, I also try to identify relevant pieces of information from some text data. And I'll show you plenty of examples for that. And then, I use a lot of machine learning. How many of you are familiar with machine learning? Little bit. So what machine learning does is it tries to learn on its own, or find on its own, patterns and regularities in a little bit of sample data that you show to the learner and use that to make decisions, predictions to new and unseen

data. Like a baby learning a language. You taught it a little bit and then you expect that it develops an understanding of how your language structures and functions, and can form a coherent new sentence, with a little bit of practice and training. That's what machine learning is. All right. So when people talk about networks of words, something that people often think of is semantic networks. And this is a classic example of [indecipherable] semantic network, and it works like this. You give people, you tell people a word or a concept, and you ask them what do you associate that with. What does that remind you of? What do you think of when you hear it? So for example, when you tell people red, something that might come up is languages, up in this cluster here, no, the colors in this cluster here, things related to fruit, things related to flowers, and so on and so forth. And then you can see there are some naturally emerging clusters which kind of get you to meaning of the concept of red. So the meaning in classic semantic network analysis-no classic semantic network theory, the meaning of the word is it's one stop environment. Whatever gets activated when you ping something in the one step environment, defines the meaning of the word, and by clustering those things you can disambiguate those different meanings. That's the classic-and this is old stuff, this is from the 1970s, from a theory called spreading [inaudible] theory. Which basically says that if you give people a prime and then whatever comes up, either they say it or you collect it from interview data or whatever they have, this is the, this is the meaning that one person associates with that term, or group or [indecipherable]. Now the world has evolved a lot since then and the way I'm or the way I'm approaching this problem goes, goes kind of like this. We have a whole bunch of text data from all sorts of sources. Can be nice, clean, well-written, used by our data where you have come coherent syntax and semantic, can be messy text messages and social media data where you have a lot of emoticons and additional things that are typed down, can be, can be email data, it can be like [indecipherable] reports, all of those things. Then what I do is I put these from some methodology, for some computational tools in order do to two things really and I'll talk about both of them. One of them is sometimes we do not have a network available. Sometimes we want to understand a network of people or of things or of knowledge or whatever, but we cannot go and ask people. In those cases me might have written documents about the situation or system we want to study that we could either read through or analyze in some automated fashion in or to construct the network data. Let me ask you, can somebody think of an example where you would want to construct network data but you cannot go and do a survey or you cannot go and do, you cannot go and do a survey. Any, any suggestions? Where might that apply? [Silence] You want to study a social, no Mark. [Laughing]

>> Mark: I love answering questions and getting them right. Animals.

>> Jana Diesner: That is a good point, animals, or, yeah, that's a good point, and people do, people study but have you talked about ecosystems or biological networks by chance? All right, but we still can observe them, that's a good point. So any idea for systems where we cannot do observations and we cannot do questionnaires and interviews? Yes.

>> Dead authors.

>> Jana Diesner: Pardon me?

>> Dead authors.

>> Jana Diesner: Dead authors which translates into anything that's not there anymore. Bankrupt companies is something that I've been doing. Things that aren't around anymore archived data, all of those things, things from the past, exactly. Something else?

>> [Low audio] networks?

>> Jana Diesner: Exactly. I've been doing that too. We call them, more modestly, covert networks where you could ask them or you don't want to for several reasons. One, it might not be very healthy. B, those people might lie anyways, or not give you the right answer. We know a little bit about the ways of how people falsify their name, or their birthdate or their license plate or whatever. Typically change the phonetic spelling or add one digit to the last digit in your social security number or something like that. But those are situations where you couldn't do that, that right. And this certain instance would be what I just told you, things that are in your mind, but they aren't written down. Cognitive models, cognitive structures, mental models, that's a certain domain where we could do an interview there but it's hard to observe right away. All right, so sometimes when we want to study historic networks or very large networks where we can't go to, we might want to do something automated with a lot of the text data, like for example, the dead authors have written, or the emails that the people in the bankrupt company have exchanged and your reports and mission statements, and often times there's a lot of text data. It's not that we don't have enough of text data, there's typically a lot of that there. Even when you collect, when you collect network data often times text data is a natural byproduct that you get along the way if you want it or not. One example would be email data, for example. So you have this [indecipherable] of emails or authors, or has anybody ever done co-citation analysis or co-publishing analysis? Right, so you have all those things who tied to or who publishes with whom, but then you also have to hold paper, for example, right? All right [indecipherable] we have all our text data, we put it through machine that gives us reliable, scalable, robust findings or network data, and those are hard things to achieve. Scalable not so much, but reliable and robust is harder to do. And then we can use it for way you want to use it. Then you have your network data. You can use it to answer some substantive questions about it [indecipherable] system to ask [indecipherable] questions, statistical properties about a network. You can feed it into a database and do a search and retrieve it later on, or you can use it for input for the computations, for example, once you have your social network you could ask what would happen if you brought in three new content editors into this group of editors in Wikipedia. Or what would happen if whatever. Whatever intervention or policy you want to design and bring into a system, you could use simulations in order to study the impact of what you are doing. But in order to do that you have to have some network data. All right. Let me give you an example. At some point I was sitting at [indecipherable] in my office and my advisor, [indecipherable] knocked on my door and said Jana, here's what you need to do. You need to develop, evaluate, apply a methodology and computational solution for mapping pretty much everything that's going on in Sudan right now, over the last kind of ten years as a network. Okay. Sudan, I had no idea, so I basically went to Google Maps, and realized it's a darn big country. It's really large. Really large, there's a lot of people, it's in northeast Africa, it's now two countries, South Sudan and Sudan, they split up as of June last year, it's not going very well. What people often know about them with respect to Sudan is the food prices, there is a lot more trouble in there. They have oil in the ground and it's sitting there and they're trying to extract it but it's not going very well for [indecipherable] reasons. Most of the oil is in the south, the south doesn't have any access to a coastline, they could go through the north but the north will charge them a lot to go through. They could build a pipeline through Uganda but they need investors from China to do that. They are religiously split. There are a lot of Arabic cultures and religions in the north and Christians and native cultures in the south and rebel groups controlling the countries, so the people in the liberation army for example, or LRA, so they have all those problems. And what we wanted to do was see who are key stakeholders, what topics or what themes are they associated with and how might this evolve into the future. So if we have a network from, let's say 2000 to 2010, can we use that to simulate what it might look like a year after. So, what I did, theory, I sat down and said okay. There are many ways to extract network data from text data. Those are they. Those are the basic families that exist for constructing relational information out of text data. If you, you can use this for creating lengthy discussions in linguistic courses. So I reviewed those methods with respects to three things. First of all, can those methods be automated? If it's green, that means yes. If it's orange it says somewhat but you have to do some manual stuff and if it says red, it's a flat out no. Second question was can we do abstractions on levels of generality? Can I look for a word and know that this word is not only word Smith but an agent, or it's not only University of Maryland but a location, or it's not only a

computer but a sort of a resource. So can we have some abstraction according to an ontology that's predefined or derived from the data and again, green, yellow, red scale, and does it generalize. Once we have a methodology or once we have a model that lets us get this data and construct this data from text data, can we expect to take the model or the methodology tool and apply it to some new data set from a different domain or different text type or different, instead of, yeah, no, different domain or different publishing data or something, can we expect it to generalize with reasonable accuracy to a new and unseen domain. And when we look at those things, the only thing that came out for me, for achieving automation abstraction generalization for our probabilistic, graphics and models, which is a subfield of machine learning and I said, okay. This is for grad students. If you really like method, make a chart of all the methods that are out there in a field, classes tie them according to some dimensions that are relevant for you and make sure that what you want to, what you want to use actually is the most suitable thing, as opposed to going to your advisor and saying we'll use semantic parameters. Somebody will ask you as some point why that method and why not centering [indecipherable] analysis or something like that. All right, so probabilistic graphical models and I will not go much into it other than telling you-I'll give you the recipe for how they work in a second, so the next thing I did was thinking about what do we want to know about in Sudan? What types of notes do we want to have in a network? What kinds of things are relevant to consider? So, for one, social agents, people and groups. Any reference to a person or to a group, you want to get from the text data. From that, you can build a classic social network right? But then you might be interested in places, references to places. We might also want to extract tasks and events and meetings and agencies providing support, people leaving for a refugee camp or coming back from a refugee camp, people building a well, all those things are tasks and events that might be very relevant for studying current issues, future developments in a country like Sudan. There are things that refer to how, to the how something is happening, resources and knowledge. What do people have at their disposal, what can they use? Natural resources, weapons, computers, whatever you have, all sorts of resources that you may have or not have. Information and knowledge, what kind of knowledge is floating around, what do people know, what [indecipherable] do they have. And then we also looked at beliefs and sentiments which is something that [indecipherable] talked about, for example, in what way is something perceived, what knowledge do people have about something, and of course we want to do this over time, so we also need to find time references. Some things might be from the past, from the very remote past or very distant past, things for now and then for example, discussions about future events like they had a peace referendum that was running out a year ago and whether they would decide whether [indecipherable] or not. So those are the source of things we want to extract from the data. [Indecipherable] people. You're not only looking at instances of people or references of people in the text data, but all those other things too. And not only that, those things can be referred to with a specific mention like Ben [indecipherable] for example, it's a specific reference, or Ben could be referred to in a very generic way, a student, a person from Pittsburg or something like that. So sometimes you have references to specific individuals and that's sociologically that's a big distinction. You might have references to specific things with a name, or to the collective, the kind of thing, to the type of thing, to the group of people. And often times, those extractors let you pull out a specific thing that's a lot easier to do because they have a different surface pattern with coupleization, all that, then finding references to collectives and more [indecipherable] preferences. But when you talk, for example, about resources, also finding things like rice and beans and public water and all of those things, might be relevant as well. So I wanted to make sure that I get the specific and generic references. And then here's basically how this is being done, and I'll try to say this in a way so that everybody thinks you can do that too, because you can. You get first some labeled data where somebody, specifically a well-paid person, sat down, it better be a well-paid person, somebody sat down and took some text data and marked all examples of what you are looking for in this data. A provider I often work with is called Linguistic Data [indecipherable]. They have very well skilled, very well trained people who provide this data with very high accuracy, so I don't have to do it, or my grad-I will not make my grad students annotate data for machine learning if I don't actually have to. [Laughing] You want to talk again about mental sanity? Right? It's possible but at the scales you need that in order to build a good model, you need a lot of this data and then you need two people to

annotate it. And then you need to measure in the coder reliability and it can take a very long time, so what I often try is to find data that somebody has already annotated. Maybe they did it in a little bit of different naming convention or something, but spend some time to look hard if you need to do supervised learning and see if somebody did the job already for you. They might have done a good job and if they did you might appreciate it too. So we have some supervised data and then we basically give this data to a tool that looks for all sorts of clues it can find about every single marked up thing. The word itself, how it is embedded in context, how it's embedded into larger collection of documents, what's right before it, what's right after it, what's surrounding it in a larger context, all of those things. And the things that I typically consider is everything [indecipherable]. So what's the word and everything related to, you know, the next word, previous word, syntax, the grammar, the structure of something, and a lot of statistical information from the data. And then you basically give this annotated mark-up data to the system, it will figure out all those features, or you tell them what clues features to use, and then it will come up with a prediction model that you can apply to new data, and let it label the new data. And this new data might have new words, new documents, all of that, that the learner has never seen, but because it can reason about every single word, given its context, and the work itself, it can tell you with some level of accuracy that we see a new thing, John Doe, and because we have observed similar things being referred to, we could say it's a person most likely, or a specific instance of a person, for example. So, I've got a thing like that. I look for all those things that I showed you, people, places, organizations, resources, knowledge, sentiment, tasks, all those things and with distinction between generic and specific references, and the thing we built has about 87 to 89 percent accuracy. Now try, try to get to 80 percent in the coder reliability. Have people done this? Have people coded stuff and measured in the coder reliability? It is not bad. 80 basically passes the [indecipherable] threshold, maybe, hopefully. I would, I would let it pass. And then you'll make this available as a tool that people can use. Actually this is available as a tool that people can use. It's part of a software called Auto Map that has those prediction models at different levels of specificity and find weakness for text data. So you can, you can use this for your work if you want to. Okay, so we have those things. So what I did was I got, I collected the data, the text-actually I was given a digitally enhanced text [indecipherable] worth about 80,000 documents about Sudan. Those documents were collected from news file sources, there were other documents going in from, there were legal documents like hearings at the International Criminal Court, there were subject matter expert reports, all of those documents together that were taken into consideration, there were about 80,000 text files, which is not a whole lot. I applied that thing to it which basically let me pull out all entities we were interested in and then the question comes up, okay, once we have identified our entities, which are the potential nodes for our network, how do we connect them? And again, there are a lot of different methods, and I'll spare you another methods review, but there are a lot of methods for how you can identify things that show up in text data. The most common and actually based on my validation, other validations, not terribly bad [indecipherable] is to use something proximity based that says if something's occur in the same sentence or in the same paragraph within a certain distance, they have something to do with each other. We don't know what, but they might share some sort of a relationship. And the results I'm showing you in here are based on that, but there are more sophisticated approaches for how to do that. We have done some of them but, and those are a list of things you could use, but what you will see is basically based on core occurrence. It's a very common method that people use when you look for things in text data and want to see how they are related. All right. So I extracted my network data and this is the result for social networks. I basically pulled out all specific agents I could find and did classic network analysis. I built a social network. In those charts you see people, they are ranked with respect to a certain metric. Decreased, you all talked about these things right? Decreased [inaudible], have you? This week? Have you talked about? Different dimensions of power and influence. So I found my top people, I ranked them over the years from 2003 to 2010, this is how long my dataset went, and they are colored according to different things. Red people are presidents or former presidents from the north. Those are the key players. And I looked at them and I marked them up. So, whatever is red, current or former presidents of the north, green are former or current president of south Sudan, orange are people you know, and some of them are not in power anymore, but those are presidents of neighboring countries.

Those are the key players that came up. And then I was looking for those things and there were some of the names, even after dealing with Sudan for a while, I had never heard. And those were things like [indecipherable] and [indecipherable] and I'm like what. But luckily for this project we had subject matter experts. Somebody who studies Sudan is called a Sudanist. So we had a group of Sudanists. I went to my Sudanists at Rhode Island college, the team was led by, their team was led by Richard Loden [assumed spelling] which was a Sudanist, he has lived in Sudan for a long time, he had studied Sudan, he's an anthropologists, and I said, Richard, who are those guys. And he said ah, those are really old religious spiritual leaders. I'm not sure they are still alive. They might be but they are really old. One of them introduced [indecipherable] for Sudan many, many, many years ago, and, but the data are from 2003 to 2010 but when we go back into the data and look at occurrences of those, or references to those names, it's often that people justify legal actions or proposals for a new constitution by referring back to the spiritual leaders that the country had. So some of those things might pop up as important entities even though they are not around anymore or they are dead authors or whatever. All right, so we did that for people and we also did it for organizations. It's the same thing. I did the same thing for organizations. I looked at all specific organizations and ran networks and [indecipherable] metrics over them and pulled out a group that I, and what you see here is everything in white are non-Sudan use groups. Everything green are people from the Sudan, actual Sudanese organizations. And something that came out was that among the top ten groups, some of them are tribes. Think of, for example, and tribes play a very central role in the [indecipherable]. Often times, they, they control all the units, but the presidents come from a certain tribe and they support their tribe and they have a lot power and influence. Often times, that's your clan, that's, often, that's your important unit of operation in certain regions. So what we did was we pulled out all the tribes and looked at, looked at tribes. First of all I pulled out all the tribes and there were 243 tribes, and those things are pretty easy to verify for a subject matter expert. So I went to, I went to my subject matter expert and we worked through the list of organizations and identified all the tribes. There were 243 in that country and they have as many languages in this country. And we looked at what topics could [indecipherable]. So that's one thing you could do once you have extracted knowledge instances and people instances, you can always look at how are different people connected? What topics bring people together, or what, what, what social clusters do we see with respect to a certain topic. So we looked at the-and those are coded on a higher level skill of topic. So we go from the word level to some higher up aggregate of topics. And you can see, for example, with respect to decrease [indecipherable] one thing you see a lot is conflict in population. When we look at between those, we see more things related to economy and land use. So that's one thing you could do. Another thing I did was for all the tribes that are either in conflict or in war with each other, and there are about 80 of them that are currently in war with each other. For all of them that are reported to be in conflict or war with each other, what resources are associated? That's a-has anybody worked in conflict management or something like that? So one thing that people often look at is when you have a conflict what sort of resources is it about, what sort of natural resources is it about, [indecipherable] or non-[indecipherable], and there are a lot of theories about it. But what we just-and for those people this is a useful methodology saying you don't need to go to Kenya, not a good example, but you might not necessarily go to a country and study all those things, but observing them or sending your very expensive anthropologist there, maybe you can do something like that and see, okay, what, what, who's in conflict with each other and what's the conflict about? So then I did this mapping, the yellow notes and the things with the yellow circle around them indicates conflict and war, and everything, the green things, are resources related to that. And there were a couple of things we could see here. And, again, interpretation was done in collaboration with my subject matter experts because I could do but it's absolutely bogus if I do it. I could look at this and say, mmm I think this is going on. But if you have some subject matter experts who you can sit down with and they can help you and also point out arrows there that's very valuable. It takes a lot of time but it's very valuable. Okay, and we saw some things, so one of the things that always stuck out are things related to Dar, that's the Arabic word for homeland, and that's the homeland of a tribe, and so one big problem that we could see from that is often times when people come back from refugee camps and they might have been there, for example at the border of [indecipherable], there's a big refugee

camp in Sudan, people were born there, people have lived there for 20 years or longer, and when they come back somebody else has already claimed the original area that those people probably had. So you have a lot of civil conflict breaking out over that. Due to climate change in south Saharan Africa and also north of that, a lot of areas that used to have water like the area around Lake Chut, are dried out and people who have herds, cattle herds who didn't move around much, are moving more into the land of farmers how always had been in, [indecipherable], or else they couldn't be there. But there were a lot of struggles between tribes that are primarily farmers versus herders and who tried to get access to water, that's a big conflict. And a lot of conflicts related to oil and buildings, transportation infrastructure for the oil. In fact when you see some of the conflict breaking out over oil, it's not actually that the oil is already there, it's often the promise of oil in the land is there and then all the struggle starts about how do we get it out and who has access rights, and who gets what sort of payment and all that. Yes.

>> Since you, maybe want to delay this the end, since you were extracting these from public reports, is there [indecipherable] you're not discovering the underlying con, the underlying reasons so much as the interpretation of the paper reporting? So in other words [indecipherable] they are imposing theories about oil and so on?

>> Jana Diesner: Oh absolutely. So does this one not match right, what's in the data, we will just cover in all the biases that are in the data will be in the results. Ways you can mitigate this by using different data sources, or you could say, given this [indecipherable] from international analysts, these are the results and what might be the biases in there, we don't know, but maybe some-and people always trust subject matter experts. I don't know, but people always do. And if you go somewhere and use and develop validation methodologically mmm, but if you say subject matter expert validated it, much better. I don't know but I see this a lot happening with reviews. So when we go to our experts and say what biases do you think are in there, they might be able to help us with it. But you're absolutely right, whatever biases and propaganda and whatnot is in the text, will be, will hopefully bubble to the surface and be visible. I agree. Yes?

>> Concerns there are different sorts of bias not in the inputs, but in the interpretations. Subject matter experts interpreting these things, do you have a sense of whether they're just saying things that thought before or whether they're actually changing their mind as a result of?

>> Jana Diesner: Oh, I will not doubt a subject matter expert explicitly. But maybe, and that's a big debate and you could come up pros and cons for either side of the debate, maybe. I mean of course you also do this with network visualizations right? You give people a picture of a network and then they will interpret all those stories into it, and some of them might just be a reaction to what they see and others might actually reveal underlying dynamics of what's really going on there. So some of them might just be reaction, other's might be fundamental cause of things, whatever. It's hard for me to separate those things. I don't know, but that danger is always there as well, especially virtualizations.

>> As the empirical question, was there any case where you showed some expert and the person said wow, I never thought of that before, but?

>> Jana Diesner: Yes, there were things like that, for example the thing with farmers and herders, they couldn't help me explain it but when I saw all those things with crops and livestock and that, it didn't make a whole lot-I mean it's a rural country right, but still why is it so prominent with respect to conflict and war. Whereas the thing with Dar's and homeland conflicts or civil conflicts about native regions, that was much more obvious to them. Any other questions right now? I like how you come up with those limitations and all of those problems, that's the job of a researcher, yes, you know, if I give this to political analyst, they're like yes, we don't need to read all the reports anymore. If I give to people like you, you're like but there's all those places, and I'm like exactly, exactly, that's important to talk about. All right, so there I am, and I did

this thing, this study and somebody says, Jana there's all the information extraction and relation extraction thing from texted, it looks like a nice thing, but how accurate are your results actually? And then some other engineer comes along and says, well, you reported this accuracy metric, it's called F metric in my domain and every domain has some sort of F metric, whatever your accuracy measurement is, this F is a combination of recall and precision, how many of the correct things do you retrieve, and out of all the things you retrieve, how many of them are actually correct and you build an average of them, and then you have your magical F measure, I don't even know what it stands for, but this is the measure we were allowing, it says 87 to 89 percent good. And then [indecipherable] comes says, but she has an F [indecipherable] so we know how good it is. And then I stand there in my little grad student office and say yeah, I tried. They both are kind of [indecipherable], maybe just the F metric is not enough of an evaluation. Maybe we need to ask some other questions. So the F metric really tells us, the F metric answers a classic computer sense question, which is how can we build a system which is this much better or this much faster than the system that a competitor has, or some baseline that's out there. But what about, what about other problems? And then I also started to ask those other questions, which go like this. How much of a difference does the selection of method for constructing that work data make, for the network we actually see? So how much of the results and he structure of the network is due to methodological choices, we as analysts make when we construct the network data? This applies to survey data as much as it does as things extracted from text data. People have studied this for survey data, biases, all those things. I did that for text data. And why does it matter? Well for one, if we can tell people 5 percent of your network or your nodes or whatever are just due to the specific method you use, then people at least know. It increases the transparency and understanding we have of a method and if you use different methods, it makes them more generalizable and comparable. So that's one of the research questions I ask a lot, the purely methodological question. And I'll show you, I'll show you where [indecipherable]. Sometimes I do studies like that where I take the same data, so we have the same dataset, and code it as a network with different methods, and then we look at how do those things compare. So here, does this mean, riding my bike, meaning things coding things by hand, it's really slow, but you see what you're doing. So I take text data and I build a code book from the text data. I use the code book, apply it to the text data and generate some network. Then I take the exact same data, use the entity extractor that I showed you, the thing that I talked about that automatically extracts those things, the little Ferrari, and it really is a little Ferrari because it goes really fast, but I think those things take a long time at applying them, it's very fast, a matter of minutes really, whereas building a code book by hand has, have people done that? Build a code book by hand? It's more a matter of weeks, or more. Okay, so we have the bicycle, we have the Ferrari, and then sometimes you could go super convenient and you could say well you have a text, smart average. You could have the text data collection and with text data, you always get a lot of metadata. So for example in the case of newspaper, think of index words that people or some algorithm assigned to every article. So you could just take the metadata and say I'm not even going to bother looking into the content, I'll just take the metadata and only keywords per article I link to each other and there is my network, in a matter of seconds, super Ferrari, high speed train. And then you could do the total opposite, this should have a snail somewhere, I will add it in. You could work with subject matter experts, which I also did. I sat down with the subject matter experts and we drew a map of all tribes in Sudan, all 248, for eight years, one map per year, and all that connections. Now this is a matter of months. Takes a long time. So we have those four different methods, but they operate on the same basis which is a corpus of text data, and if you complete it well, they will give you vastly different networks, but we don't know how different they are, but now we do, now we do. And I did this for, and I'll speak three more minutes. I did this for three different datasets. I did the same four methods on all three datasets. One of them is the Sudan corpus that I showed you. The second one is a corpus that I collected from something called Cordis which is a database of all of proposals that a European union has supported with research money since the 1980s. That's 20 years of data and it basically has for every proposal that got funded. I don't have the proposals that did not get funded, but for those that got funded, we know who proposed it, who are the collaborators, all those metadata, how much money was awarded, all the keywords, and then we have the proposal [indecipherable] right, the actual

text stuff of the proposal. So we have the text portion and the social network portion. Who works with whom and what was the proposal about. And then I also used my version of [indecipherable] data, where you can do the same. Who sends an email to whom and then you have the body of the email data, okay? So those things are very comparable and that they are long term, large scale data, very different domain, nice clean news wire data, scientific writing which is also pretty clean and then, you know, messy email data. And what we found across all, I'll make this short actually, I'll make this short. What we found, first of all one big question is, when we use any method for automated network data construction from text data, that's a resemble from truth. That's to resemble the things I constructed with my subject matter experts. If it does, we can skip the lengthy work of subject matter experts, if it doesn't, we might still want to keep them around or antiquate them in some smart way. And the answer is, not really resembled, but when we do the hand coded text book, I got 50 percent of the nodes and 20 percent of the links that were also in the subject matter experts. With the fully automated thing, even less, and the only reason why the [indecipherable] between the manual code book and the thing with the subject matter experts was kind of high was because in the end we were, when we were done with subject matter experts, I also told them of tribes that we identified and put them in my codebook, because this is how you develop a code book, by working from some [indecipherable] data and expertise. But even then, the network overlap is pretty small. And when we construct, of you want to read this my dissertation was about it so you, and there's a paper too. When we construct different types of networks, the social network of who talks to who versus we look into the semantic network of the text bodies or the proposal data or the email bodies associated with it, it's very different. To give us different views on the network, the social network, the text based, the text based social network gives you a pretty good view into localized agents. There's a lot of references and mentions of people on the ground and the look of people. Whereas if you just take metadata, you know the thing that I just constructed from entries in a database, it gives you a high level view of the major international players, but hardly any people from the specific country. It's not really there, at least with the lexis nexus data we used, and there was a lot of lexis nexus data. So, it's not really there. The [indecipherable] and we looked into the bodies, the knowledge networks, semantic networks, things constructed from text bodies, and things constructed from metadata seem much more informative and that's because the key indicators, the keywords that you assign to a proposal when you submit it, for example, are very good mini summaries of the data or of the text that you already actually have in there. It's good descriptors. They read well, they are nice good keywords. Whereas if you knowledge networks you get more, you get more broad concepts like research, science, methods, all of those things, but it's not very specific to the content domain. It's like what social networks do for text based in terms of assuming it into details in culture, the text bodies do format data networks. And what do we learn from that? So there's a happy ending to the story. So we will marry the Ferrari to the metadata. Those are both very fast methods. We use text, we look into the substance of text data by basically using automated entity detection and relation extraction, and combine that with the metadata network because those two things together give you both things you want. They give you a high level overview, a very general bird's eye view of the key players, but also some deep flavor of the culture and more rich things that are going on in a domain. But you might see here that the actual [indecipherable] this can be replaced by combining these two methods together, which for me, is a very happy ending of this research. Okay, that's basically what I-and I would have shown you more if we got bored really fast, but we didn't. And to be a good citizen here I quickly want to say, well first of all, you know, there are always sponsors happy. And then do how does this tie to, and you can see how this ties basically to all the goals that were set up for this week, basically, so applying those methods, well how does this help to clarify national priorities, using those methods to analyze large collections of inherently qualitative data in a sense, in a large scale, can help us to scale up analysis and typically anthropologists and sociologists did by close reading methods, to a much broader scale. We have all these deep, we have all this rich qualitative data, but before we would, you know, pick a few data points and then assume it. Now we can also look at them on a more bird's eye view and then zoom in, but we can analyze them more quickly, more efficiently on a larger scale. And looking not only into, you know, your common entities, people, places, organizations, but all those additional cultural references, all resources, all the knowledge, all the

sentiment, all of that, on a large scale. All right. So, with that, if you have any questions or if something in here seems debatable to you, which I hope it does, or if you have any follow up questions let me know, I'd be happy to hear about it. All right. [Applause] Any questions right now? Nobody feel asleep, I am so proud. [Laughing]. Here you go.

>> I was wondering how much checking of the metadata you use before you use it in the computational methods?

>> Jana Diesner: What do you mean by checking?

>> Like some of my friends have tried to use Google Books as a source of the data and they have to approve and code some of the metadata because it's so bad.

>> Jana Diesner: Yes, see this is because they are scientists, but talk to some industry people. They rely on that. They take metadata, and if you look into it metadata also has these ambiguation issues it's not that somebody at lexis nexus sat there and took all the different ways you can say [indecipherable] which is the president of north Sudan, there are many ways how you can shorten this version to just [indecipherable] and then there are a lot of versions how you can actually spell Mohammad in English translation. The metadata has less of noisiness than pure text data, but it's not perfect, but people often think it's perfect. In the paper I have about this, it actually says about how perfect it's not. It's pretty good, but it's not perfect. I wouldn't even say it's good enough, but with a little bit of entity resolution, this ambiguation you get to put enough very quickly actually. It's not a nightmare. Other questions? Comments? You guys want to take a break? I guess we ended right on time.

>> Are your papers up on a website?

>> Jana Diesner: On my webpage there is, there is, I can show you that. I guess I'm not online right now. If you Google my name, there's basically only one Jana Diesner in the world, and it gets you right to my webpage and on my webpage is a publication section and it has those things there. Thank you very much. [Applause]