# Interview: Participant 2

# I: Just more like a general chat but I have prompt things if you don't cover it anyway. So first off, could you tell me your discipline and your research field in your current role?

F: Ok, so I'm in Geography and Environment, I'm a transport researcher, technically. So that sort of encompasses a broad range of things that I study and things that I work on**.**

# I: So quite a range. Just very brief, but could you give me an overview of what your research is?

F: So, it's broadly the collection, use and communication of data for various purposes, predominantly related to transport and transport efficiency so it's basically looking at how can we take data that are being generated either purposefully, so data that are generated, for example, through something like Twitter. Or data that is created for purposes but not necessarily for the user, so things like GPS data from phones, location traces, that sort of thing and how those can be taken together and used to provide more efficient practices for transport.

# I: So how do you currently use social media in your research?

F: A couple of different ways.. A set of identifiers so particularly things like Twitter users in the area, key words that might be relevant to whatever, so we collected data to look at how information is disseminated during large events using social media as a platform. So whether it was just general information whether it was information about transport disruption, whether it was – and people looking for information or providing information, that sort of thing. That's the primary one, now that's also linked with – there's a lot of overlap between that project I work on that, or work on both of those. So also looking at how the operators use social media to share that information and how people respond to it. So we've done some other – I also do a project using Facebook data and dealing with social network analysis of pages and the relationships between pages and Twitter, so looking at different matrix, closeness between the different centrality measures, density measures, that sort of thing. So we use that to try to give them some familiarity with the different type of data set and also with the different type of statistics.

# I: So the next one was: What social media do you use, the likes of Facebook, Twitter, Instagram, do you use in association with research and also for research?

F: I'm sad – when it comes to social media so I'm terrible at participating in it, I look at other people's stuff, so we use it a lot for research but I am abysmal at sharing information on social media!

# I: That is the same as me! So, we've discussed it a bit already, but do you use any tools? And by tools I mean anything – so you've said you've got the Twitter monitoring interface…?

F: Yes, we use data collection and we also use things – like Note Excel and Gephi, which are the three softwares, do you know them?

# I: No, I've never…

F: So Note Excel, it's a free open-sourced software that's based on an Excel platform so you basically install it and when you open it up it just comes up like an Excel spreadsheet and there's an extra little area and you can use that to import data, to classify, categorise, visualise certain bits of data and it's a handy little tool actually. And then Gephi is – it's basically, it's a free open-sourced visualisation software so you can import things like a Twitter network or Facebook data, anything that's sort of structured data that's set within a network and you can use Gaffe to analyse that. So you can do – again, it does the visualisation, it can also run the statistics in there.

# I: Oh, that's handy.

F: Actually it's really nice and it's on many of computers because I use it for ut this on the computer

# I: That would be good. The next one was about how do you store your data when you collect it?

F: It depends on the data. So most of the data that we use are public data so it's information like Twitter – we only get the information that's been made publically available on Twitter and again with Facebook, they recently changed their privacy policy such that you are no longer allowed to export data on a personal network but you can export data on a page network. So with that, what we do is just – that's an example of the visualisation. (…) I train it so this is…

# I: This is all the stuff that comes out from peak fitness first.

F: Yes, so what it does is it looks at the likes, so if peak fitness has liked a page or if that page has liked peak fitness then it will come up so this is the first-degree *(ego-network*) so this is just ones that have an immediate connection so either. You then go into a second degree network that then pulls off other pages that have been liked by that group. It's not the best visualisation but sort of an initial one. Yes, so when it comes to storage, most of it is just stored on my drive so it's a small enough size and has a little enough personal identifying information that it doesn't have to be *(locked) down***.**

# I: So do you literally just take a copy of the Twitter data at that point in time?

F: Yeah.

# I: Do you ever review whether or not it's changed because you know how people can, in theory, go back and delete or change…?

F: I have at times. What I try to do is every time we run a new analysis, so try to use a fresh set of data, so I keep the same rules and then I set up – because we mostly use Excel for the analysis, we've made the – we cut them into small enough sub-sets that they can be manipulated within Excel. So what I try to do is set everything up there so it can be copied over. So yeah, we do try to keep in mind that these Tweets can be deleted and certainly the *(00:09:00 total-monitoring)* infrastructure takes that into account.

# I: So if the Tweet was deleted would you remove it from your data-set or would you want to know that it had been deleted in terms of…?

F: We would need to, part of the Twitter user requirements, we would have to delete that so technically – you can have a static-set and I don’t know if you are expected to remove it from that but the Twitter monitoring infrastructure has to be set up such that if a user deletes the Tweet it's removed from our…again, that's per Twitter.

# I: So we've discussed it a bit already but for analysing, is it mainly Excel you use and then..?

F: Yes, it kind of depends. So we've done, for example, with the Twitter and project, we've been looking at different categories of Tweets so things like information provision, information-seeking – what type of *(00:09:52*) is the user participating in when they communicate with the (*populator*) or what type of information is the operator trying to share with the users. So we went through and we did some categorisation and classification exercises. predominantly in-between myself and we did – we each ran categorisations and we came together to (generate) the reliability tests and then did another – we did, as an *(interim processing tool, we reached a crux of agreement* and that was all done in Excel. Again, we've used things like Gephi for the visualisations – this is what this particular Twitter network looks like and it's nice that you can – particularly if you have a public presentation if you can do this short analysis using something like Gephi so it gives you a nice picture but also some nice stats in the background.

# I: That's good, yeah.

F: And so the categorisation stuff that we do in Excel is more about the content of the Tweets whereas the Gephi stuff is more the network itself.

# I: So when you collect one person's Tweets, obviously because you've been doing it a lot about transport, would you then want to know who had re-tweeted that, who had commented on it and if that network – do you want to know more about the person who did the Tweet or do you just take their username and leave it at that?

F: It depends. We gather – the information that we gather is the Tweet, the username, if they have any affiliated geographical information because we are interested in the geographic bounds of the network. We've also gone to the pages – so if somebody has a public profile we can go there and determine what type of user it is, is it someone who is a public figure, is it someone who is just an individual, is it a company, is it another transport service provider. So we do try to get some information about that to see who is actually interested, is it just the media who is re-tweeting the stuff because they need to inform their users or are other people stepping into the frame and saying, 'hey, this is something that people in my network might need to know.'

# I: That makes sense. Is there anything that you can't do at the minute that you would want to do?

F: It would be nice if we had – basically the biggest barrier to some of the analysis that we want to do at the moment, it's just time. The process of coding these Tweets can take a very long time because we're doing it manually at the moment. So it would be useful to have something – there are automated tools that you can train on a certain set of data and we've used some of that for (*sentiment)* analysis, we've used another open-sourced tool called (*Rapidminer)* to do some initial sentiment analysis of the Tweets.

# I: Because you are dealing with millions of Tweets, aren't you?

F: We've collected millions of Tweets, with every one of these we've used a smaller set simply because of constraints on time and on the programmes that we have. Because we don't have – I know there's stuff that can be done in NVivo but I'm not as familiar with that, we don't have any specifics for content analysis tools that would be applicable at the moment, at least that I'm aware of. So some more *(counter*) analysis as well so that we can develop some training data and then be able to use a larger…

# I: …data-set, yeah. Do you add any other data to your information? So you collect all these Tweets, do you use other data when you are analysing it or do you look at other information from other places?

F: We will sometimes. So obviously, again, sometimes we'll go to the public profile and find out what type of user they are. We've also talked a little bit about exploring, linking things like what was happening at the Games during that time, were there certain events that were going on, what does the *(closure*) look like, what are these patterns? So to some extent trying to link what was happening in the social network world with what was happening in the physical world.

# I: Ok, that makes sense. What would you like to be able to do with social media data and any other data sources that you can't currently at the minute?

F: Wantonly violate people's privacy. I think it would be – it would be nice to have, it would just be nice to have some more time to dedicate to it.

# I: Because you work on such a large data-set it is difficult to really..

F: Yes and I think the data-set that we collected is one that is really rich so one issue is that we've sort of taken this particular approach of looking at how data gets re-tweeted, how this might be an indication of trust in certain users etc. That's one way of cutting the data. But there's so many other ways of segregating it, of cutting it, of identifying things that might be interesting. I think that's just a big part of the (…) is – it would be nice to sort of be able to say, 'these are the most applicable questions, the most interesting questions and let's actually take some time with these sixteen-line Tweets that we've collected, go through and figure out how many of them are spam, how do we determine how many of them are spam, how do we better classify or categorise things such that we're getting a higher percentage of useful and useable data. So in some ways it's – what would be useful would be to develop better methods in the background in terms of collecting and cleaning, that's something that would be really nice. Again, my background is privacy research, it would be nice to do a bit more research around what people are sharing, how sensitive some data may be that people are putting out there anyway.

# I: The last question really was if you had a dream research tool for doing stuff around social media research, what would it do or how would it work? Because you've spoken a lot about using multiple tools and all the different features?

F: Well I think anything that, in a way – my dream tool for social media research would need to have functions like a Swiss army knife so that you can sort of – modular tools that you can say 'this is something that is applicable, this is something that is applicable – and this is something that's applicable to this particular research and I'm going to put those together. We want something that does – for this study we want to visualise the network rather than the content so this visualisation tool, this information about how to analyse that network in terms of (*node strength*) and degrees and shortest paths and how do you get from this point to this point. But on the other hand there might be times when it's like, 'ok, let's shelve that, those are useful things but what I want for this analysis is the content analysis tool and how to determine the matrix for that.' So I think for me it's taking all of those things and then being able to select which ones you want.

# I: So rather than having to come out of one and go into another and then come out of one and then do another, often you have to re-shape the data for each…

F: Exactly, exactly. Something where you can have associated files and associated information but again, there's different ways of cutting it.

# I: That's really helpful, thank you. That's really all I had unless you had anything else to add about social media?

F: No, I think that's it.

**END**