Social Networking Special Ops: Extending data visualization tools for faster pwnage

Chris Sumner
@TheSuggmeister
www.securityg33k.com

Latest Document

The latest revision of this paper will be available at http://www.securityg33k.com/wp/BH10.pdf after the conference.

This is revision **r1.2**

Disclaimer

I am not writing on behalf of my employer. The information and perspectives I present are personal and do not represent those of my employer.

Acknowledgements

Roelof Temmingh, Andrew MacPherson, Dominic White, Adrian Mahieu, Tony Hawk, Jerome Case, @l0sthighway, @alien8.

About the Author

Chris "@TheSuggmeister" Sumner has been directly involved in Corporate Information Security since 1999 and has maintained a passion for security since seeing Wargames when it first came out. After a lengthy stint as a Pivot Chart creating, PowerPoint wielding, Security Manager for a business division that alone would make the Fortune100, he has turned his attention to a more geeky pursuit and is currently focused on Security in the Development Lifecycle.

Outside the corporate world Chris is a data mining, analysis and visualization geek at heart and also enjoys hiding skateboards in the UK for Tony Hawks twitter hunts.

Social Networking Special Ops: Extending data visualization tools for faster Pwnage

Chris Sumner
www.securityg33k.com | TheSuggmeister@gmail.com | twitter.com/TheSuggmeister

Abstract

This paper describes how data visualization tools can be extended to speed up the analysis of social networks. It shows how a combination of data mining, named entity recognition and visualization can quickly draw attention to interesting social relationships. Two cases studies describe these techniques in the context of social networking.

The first case study outlines how an analysis of skateboard legend Tony Hawk's twitter hunt had an unexpected benefit of uncovering top talkers, including a member of Tony's staff..

The second case study illustrates how these techniques have been used to enumerate a 419 scam, infiltrate the scammers social network and expose deeper, more sinister links to organized crime.

The focus is specifically on Twitter and Facebook, using the Twitter API's and publicly available profiles.

Keywords: social networks, visualization, data mining, Maltego.

1. Introduction

Social network analysis is not new, perhaps unsurprisingly, its origins can be traced back to the ancient Greeks[1]. However, it wasn't until the advent of Gestalt psychology in the late 1800's that the study of social networks started receiving formal scientific exploration, most notably from Jacob Moreno, who is widely credited as one of the founders of social network analysis[2] and the creator of the sociogram[3] *(See Figure 1.1)*.

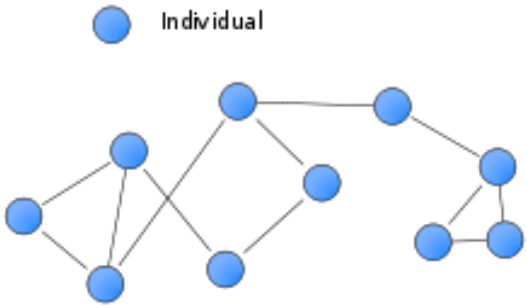

Figure 1.1: *Example Sociogram[4]*

As social network analysis is not new, neither is Social network visualization. However the majority of approaches to date have tended to follow an approach of piping the output of data mining into a visualization engine such as Vizster[5] and UCINET[6] *(see Figure 1.2)*.

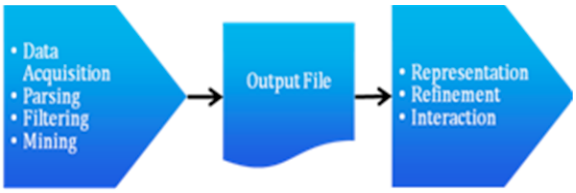

Figure 1.2: *Typical approach for social network data visualization*

While this technique is appropriate for representing data in a visual format, it is not interactive. A suitable analogy of the limitations of this approach is creating a visual image of a directory structure, but not allowing a subsequent operation on the output (e.g. file and folder operations).

Interactive data visualization bridges this gap, allowing a user to perform an action on a node, manipulate the results and perform subsequent actions in an intuitive manner.

2. Target Rich Environment

2.1 A perfect storm

Within the last decade, three key events have converged to create a perfect storm or a "target rich environment".

- Significant growth of data
- Increased use of social networking
- Increase online promiscuity.

2.1.1 Significant Growth of Data

According to figures from Cisco[7], monthly internet traffic has grown from 5 exabytes per month in 2007 to 21 exabytes per month in 2010 and is expected to reach 56 exabytes per month in 2013.

A 2007 study by the IDC[8] offers a slightly different perspective looking at the growth of online content ("information that is either created or captured in digital form and then replicated"), predicting an increase from 161 exabytes in 2006 to 988 exabytes in 2010.

What is clear is that there is an enormous and growing amount of data available and a significant percentage of this data is personal information, such as photos and videos.

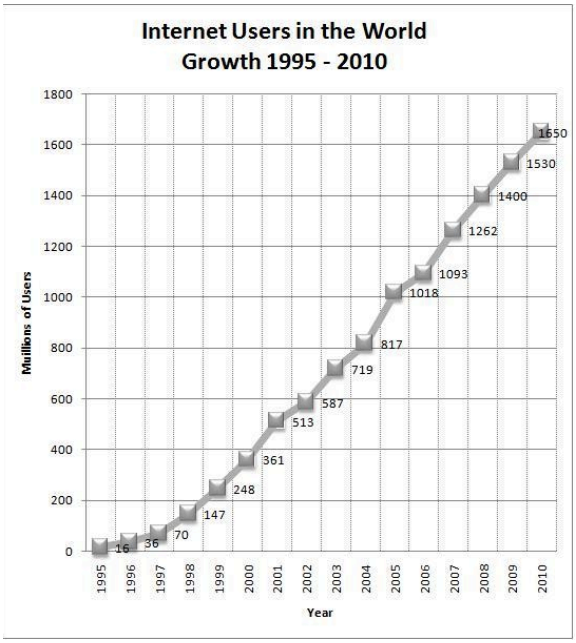

Figure 2.1: Internet Users in the World, Growth 1995 - 2010. Source: http://www.internetworldstats.com/

2.1.2 Increased use of Social Networking

If we discount email, in 2003, social network usage was relatively obscure. However, in 2010 social network usage is prevalent and isn't limited to stereotypical "geeks".

- Facebook - 350 million users[9]
- Twitter - 100 million users[10]
- Myspace - 113 million users[11]
- Bebo - 12.6 million users[12]
- Linkedin - 70 million users[13]
- Friendster - 115 million users[14]

2.1.3 Increased online promiscuity

"Online promiscuity", refers to the practice of people putting more and more of their personal information online.

Perhaps the most noteworthy result from available research is a privacy paradox. Social network users appear to be stating that they take privacy seriously[15] , yet these concerns are not reflected in their online profile settings[16]. So they are saying one thing and doing another.

From the 2005 paper "Information Revelation and Privacy in Online Social Networks (The Facebook case):

- 89% of users use their real names
- 61% of user use an identifiable image of themselves

Note: Whilst dated, these findings were corroborated in the September 2007 paper "Student Awareness of the Privacy Implications When Using Facebook"[17] and there is little to suggest anything has changed radically in the last 3 years.

It can be argued that this state has been reached through a pervasive "I've got nothing to hide" mentality. As Daniel J Solove states "In many instances, privacy is threatened not by singular egregious acts, but by a slow series of relatively minor acts"[18] . i.e. A sequence of events led to the implicit trust that many users display in sharing personal information on social networks.

For this generation at least, it is likely to be too late to turn back the clock with regards to privacy.

2.2 Why does this perfect storm represent a problem?

Data alone may or may not be compelling, but when aggregated it can expose previously "hidden" information. Perhaps the best example of this was the 2006 incident when AOL release a text file containing search keywords of over 657,000 users. The New York times [19] selected a particular user (AOL user "4417749") and extracted all associated search terms, leading them to a Ms. Thelma Arnold from Lilburn, Georgia, USA. Search mirrors [20]still exist for the curious.

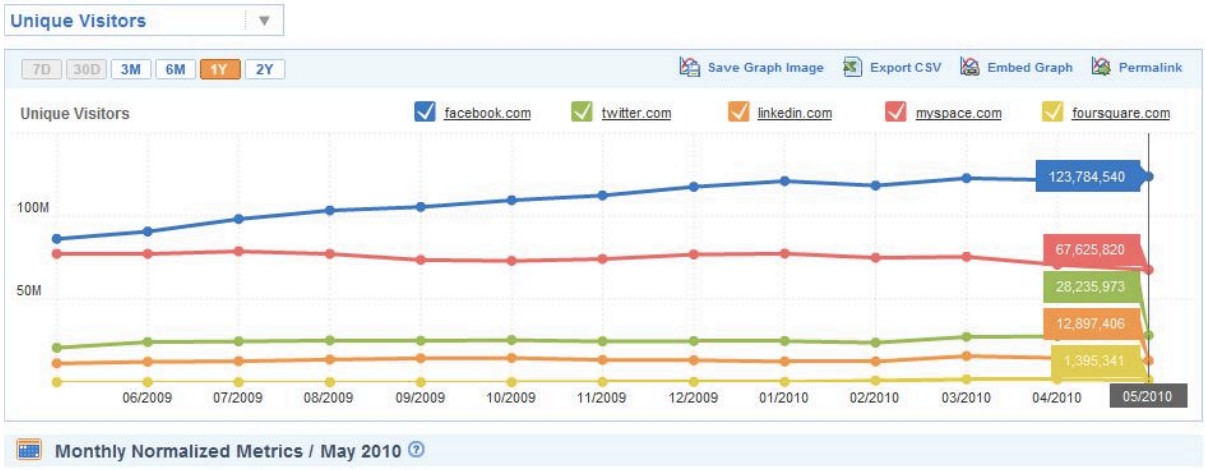

Figure 2.2 Social Network Site Unique Visitors June 2009 to June 2010. source: http://compete.com/

3. Opportunity

Clearly there is a rich vein of information in all this data, but for casual observers, it's simply lost in a sea of noise. The following subsections describe how a combination of Interactive Data Visualization and Named Entity Recognition (NER) can greatly aid the analysis of data sets.

3.1 Visual Data Analysis

Visual Data Analysis is the process of graphically representing data (potentially huge amounts of data) to highlight specific information and facilitate quicker decision making.

"The basic idea here, is that you'll notice things visually that you wouldn't be able to even see otherwise." Jim Andrus[21]

3.1.1 Visual analysis, you'll either love it or you'll hate it

In terms of how people prefer to receive information, Fleming[22] describes types of learning modes *(often referred to as V.A.R.K.)*.

- Visual Learners – a preference for visual representation.
- Auditory – speaking/listening
- Reading/Writing
- Kinesthetic – touch/feel

It is important to note that people are not the same, which often creates problems. Perhaps you lean to visual but need to present to an auditory audience. Conversely, Auditory learners can become exasperated when there visual counterparts seem unable to retain spoken instructions, such as directions.

Visual learners account for roughly 60% - 65%[23] of the population so clearly data visualization has an enormous target market.

3.1.2 Current approaches to data analysis and visualization

Approaches to data analysis and visualization broadly follow the approach of acquiring and cleaning a data set, performing analysis and displaying the results.

Ben Fry identifies 7 steps in data visualization in his book "Visualizing Data"[24].

1. Acquire
2. Parse
3. Filter
4. Mine
5. Represent
6. Refine
7. Interact

It is noted that the sequence and selection of these steps can vary dependent on the application/problem the researcher is trying to solve.

3.2 Named Entity Recognition

Named entity recognition (NER), also referred to as entity identification/extraction[25] is the process of parsing data to extract and classify information.

From wikipedia (as of 17-June-2010)

> *"Most research on NER systems has been structured as taking an unannotated block of text, such as this one:*
>
> *Jim bought 300 shares of Acme Corp. in 2006.*
>
> *And producing an annotated block of text, such as this one:*
>
> *<ENAMEX TYPE="PERSON">Jim</ENAMEX> bought <NUMEX TYPE="QUANTITY">300</NUMEX> shares of <ENAMEX TYPE="ORGANIZATION">Acme Corp.</ENAMEX> in <TIMEX TYPE="DATE">2006</TIMEX>".*

3.3 Interactive Visual Data Analysis

Interactive Visual Data Analysis combines data visualization with the ability to perform an operation on the data.

"An interaction technique is the fusion of input and output, consisting of all software and hardware elements, that provides a way for the user to accomplish a task"[26].

Making the representation of data interactive enables a researcher to quickly explore interesting nodes and relationships within the graphical environment, instantly "seeing" results.

4. Tools

4.1 Table of Tools

The following table presents a non-exhaustive list of tools, highlighting which perform visualization, interactive visualization or Named Entity Recognition.

	Visualization	Interactive Visualization	NER
Processing	Y	Y	N
Graphviz	Y	Y	N
OpenCalais	N	N	Y
Maltego	Y	Y	Y
touchgraph	Y	Y	N
mindraider	Y	Y	N
Vizster	Y	N	N

Table 1.1 Non-exhaustive comparison of data visualization tools

4.2 Maltego

Maltego allows the user to combine a variety of data mining tasks, including 'Named Entity Recognition' in an interactive visual context.

"Maltego is an open source intelligence and forensics application. It will offer you timous mining and gathering of information as well as the representation of this information in a easy to understand format."[27]

Arguably its most compelling feature the the visual social network analyst is it's 'Local Transforms' feature:

"Local Transforms are just that, transforms that run locally (the same PC that Maltego is running on). These are applications that when called will produce output which results in entities within your graph. They can be coded in practically anything as long as they stick to the specification."[28]

This gives the user the ability to mine and graph virtually any data source. This tool is therefore readily extendable to perform social network analysis through the API's often present with social networks such as Twitter and Facebook.

5. Case Study # 1

5.1 Background

In 2009 and 2010, US skateboarder and sport hall of famer, Tony Hawk, used Twitter to run a world wide treasure hunt (twitter hunt). Tony enlisted some of his twitter followers *(tweeps)* to hide packages containing skateboards and other goodies in cities and towns around the world. He then tweeted clues, which his remaining tweeps used to hunt down the packages

5.2 Objective

Create a Google map of the locations that #THTH *(Tony Hawk Twitter Hunt)* packages were hidden in, who hid them, who found them and any pictures associated with the hide and the find

5.2.1 Hypothesis

People who hid packages would tweet the finders of the packages with a note of congratulation.

5.2.2 Null Hypothesis

People who hid packages would NOT tweet the finders of the packages with a note of congratulation.

5.2.3 Approach

Associate the finders of skateboards with the hiders of the skateboard and determine geographical location of the hider and finder.

5.3 Starting Assumptions:

1. People who hid packages were all 'followers' of @HidingIt.
2. Tony tweeted each find and each of those tweets contained the word 'Found' and typically where it was located.
3. Finally, everyone was encouraged to use the twitter hashtag #THTH when communicating on the topic.

5.4 Deriving a Starting Node

Note: In v2.x of the product, we have to derive a starting twitter node (AffiliationTwitter). i.e. We can't just plonk @HidingIt onto our map and have Maltego figure out that it's an AffiliationTwitter. I achieved this by following these steps

1. Drag the Phrase icon from the left hand menu and dropping it in your main map window
2. Double click where you see the text 'Phrase'
3. Start typing the phrase you wish to search for,
   in this case "@HidingIt".

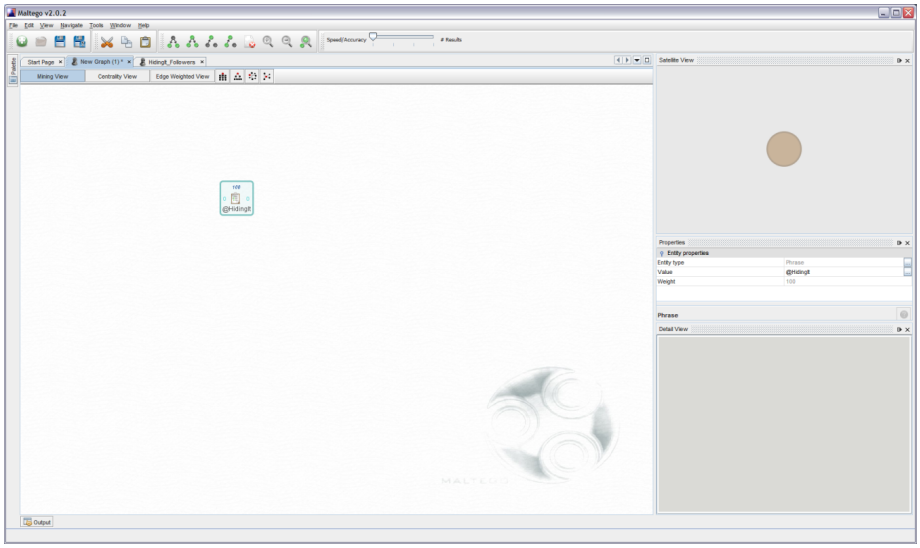

Figure 5.1: Putting phrase on the map

4. Next, find all the tweets which contain the text "@hidingIt" *(Figure 5.2)*. We do this because we'll later be able to derive a twitter entity or "AffiliationTwitter" from a tweet *(Figure 5.3)*

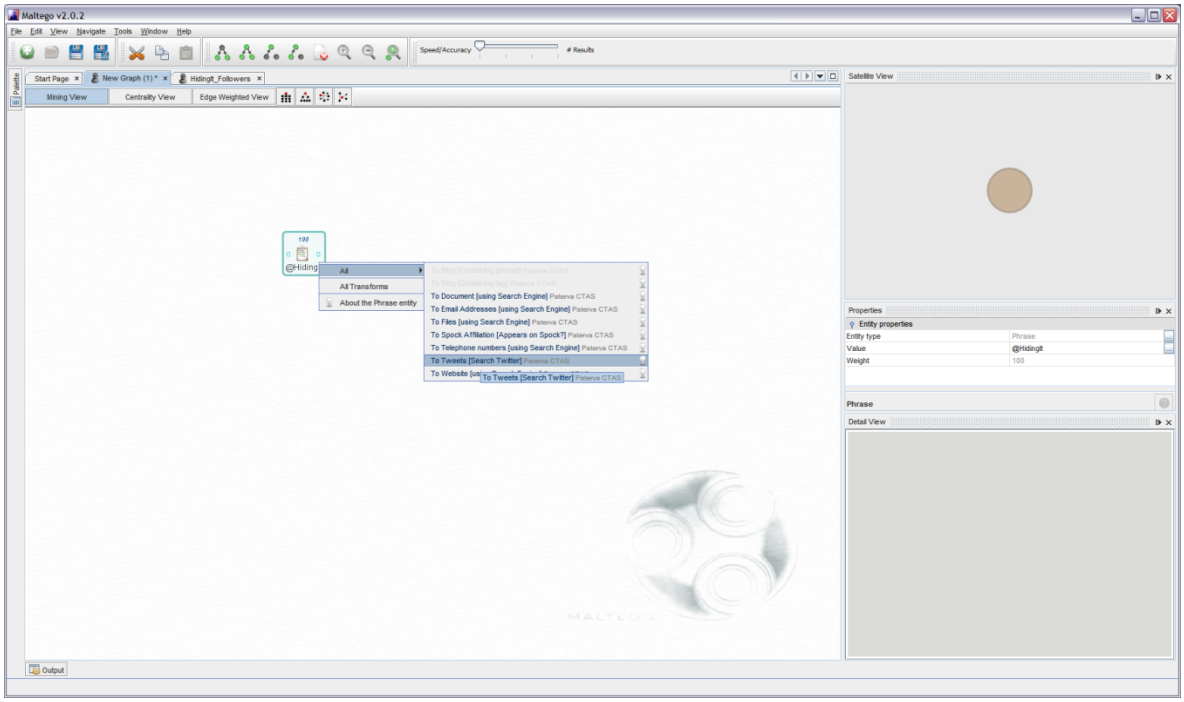

Figure 5.2 Searching for tweets containing "@HidingIt"

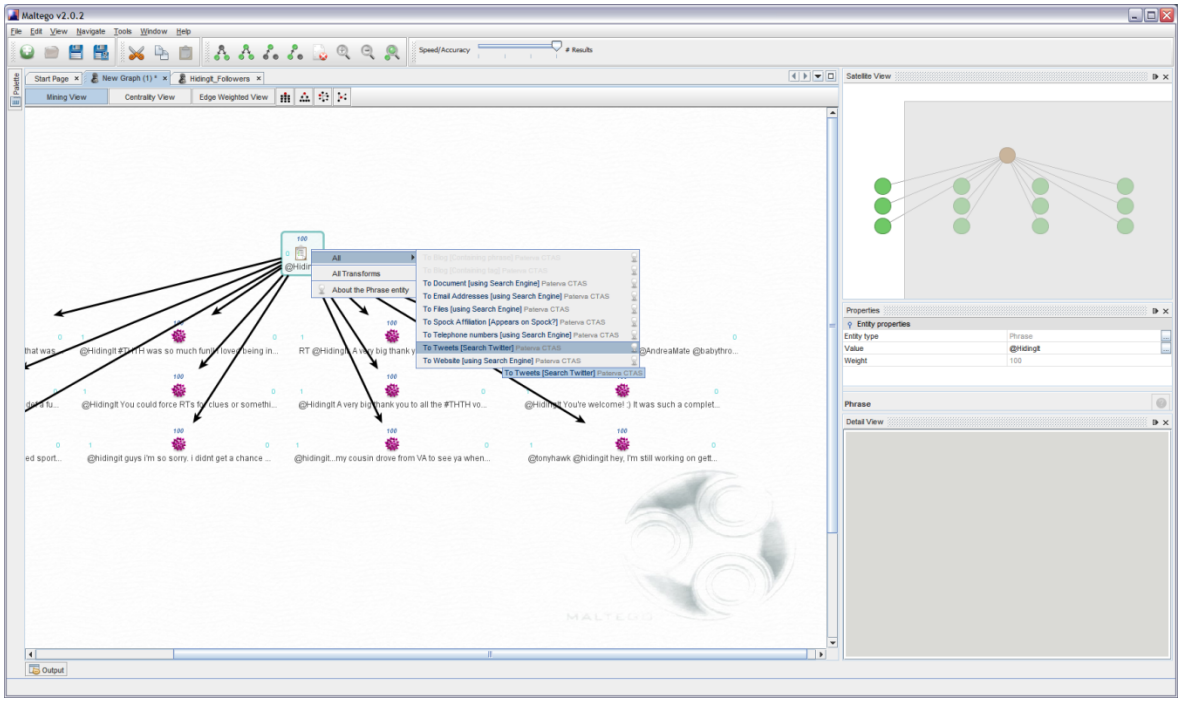

Figure 5.3 Searching for tweets containing "@HidingIt"

5. Now we can generate an AffiliationTwitter (*Figure 5.5*) by running a "convert to Affiliation Twitter" transform.

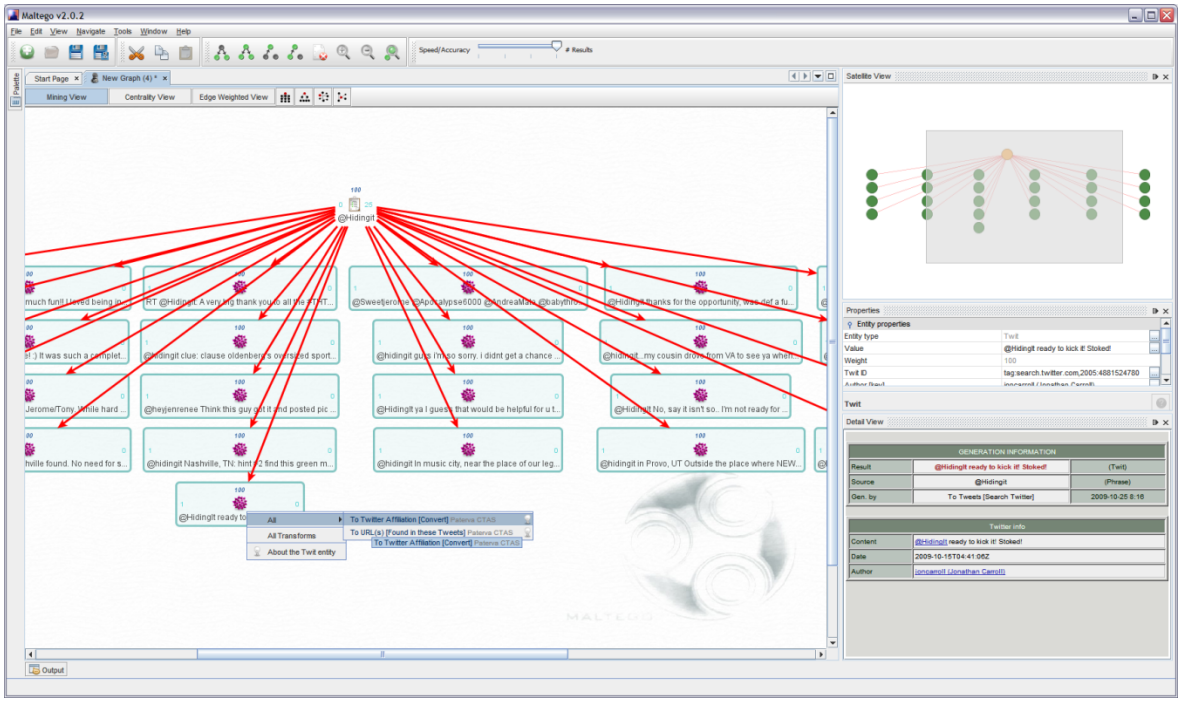

Figure 5.4 Converting a tweet to the Twitter user who sent it.

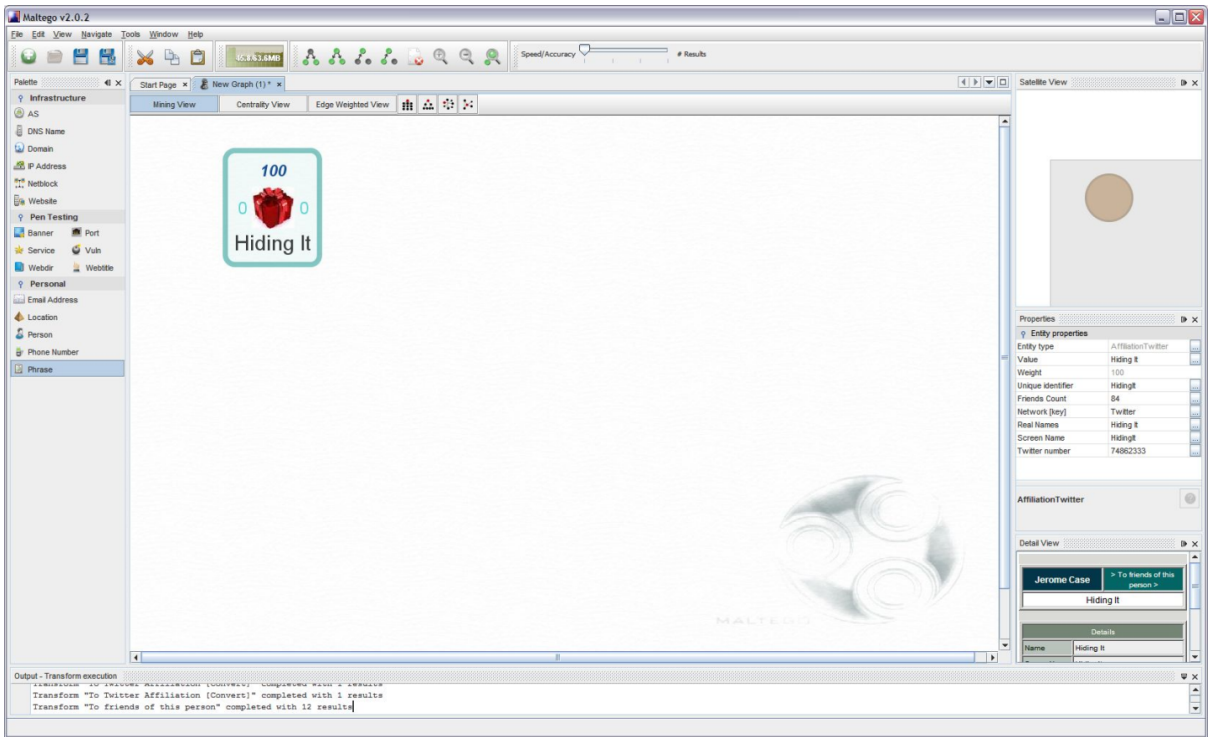

Figure 5.5 An "Affliation Twitter"

5.5 Data Acquisition:

Maltego has built in Twitter transforms, but as of version 2.02 a number of them suffer from problems caused by known limitations of the Twitter Search API. To acquire data from Twitter, I therefore constructed local transforms to use the Twitter REST API's.

5.5.1 Pseudo code

- Get followers of @hidingit
- Get all tweets written by USER within between <Start Date> and <enddate> & extract the @username

6. Get followers of @hidingit

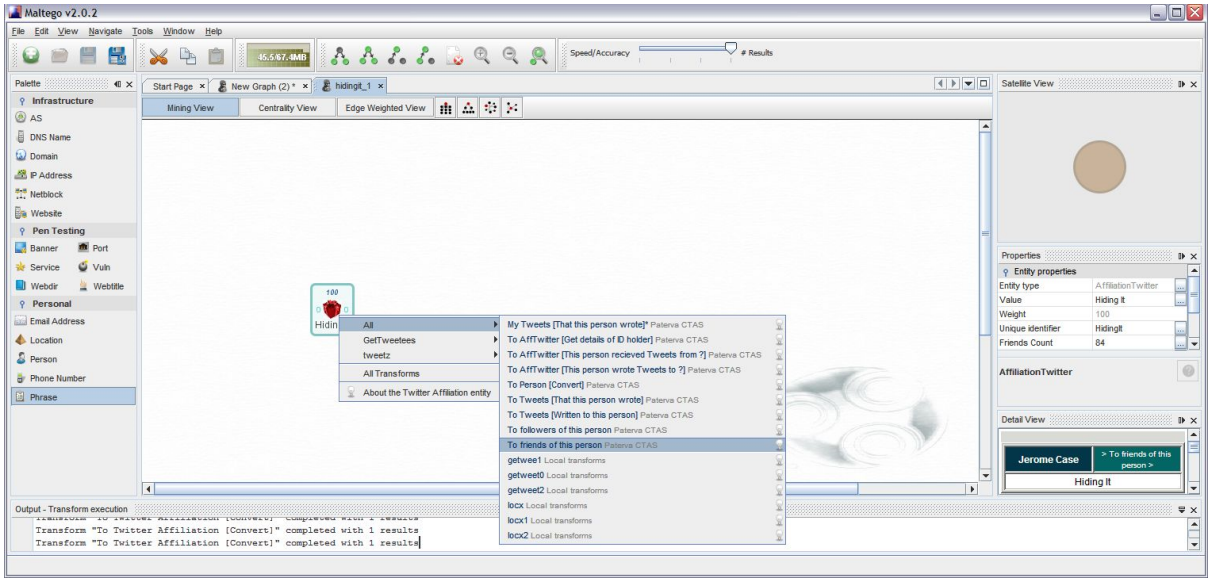

Figure 5.6 Getting the followers of @HidingIt

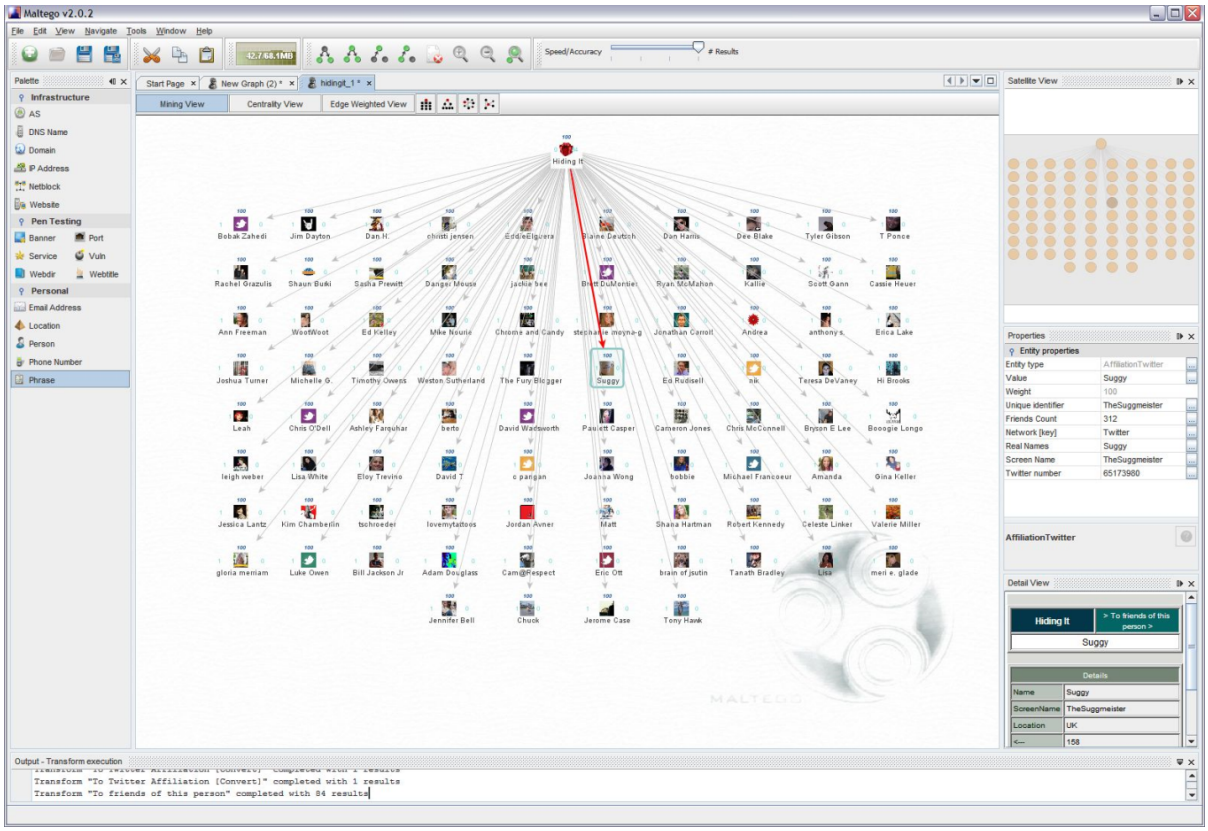

Figure 5.7 Followers of @HidingIt

7. Select @Tonyhawk (since we want to grab his tweets)

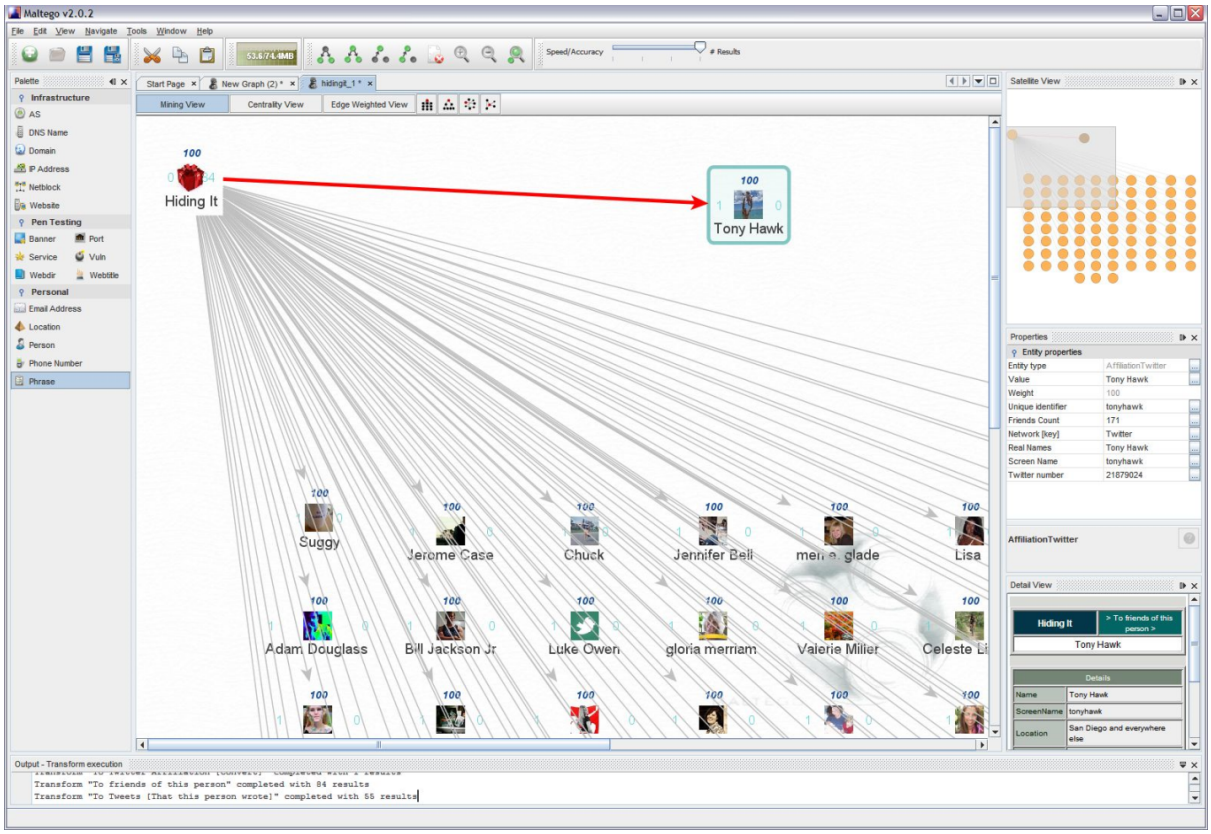

Figure 5.8 Isolating/Selecting @Tonyhawk

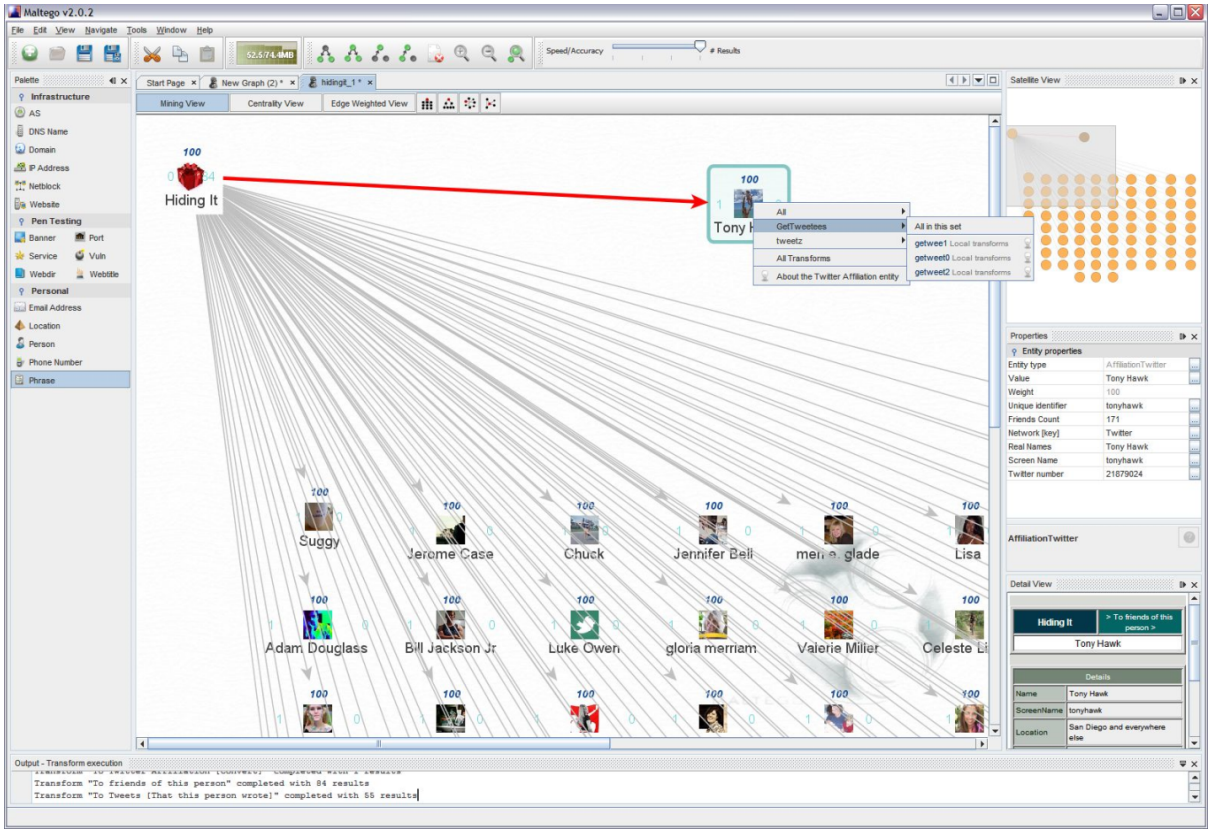

Figure 5.9 Getting tweets written by @Tonyhawk

8. Get all tweets written by @tony hawk and Extract the @username

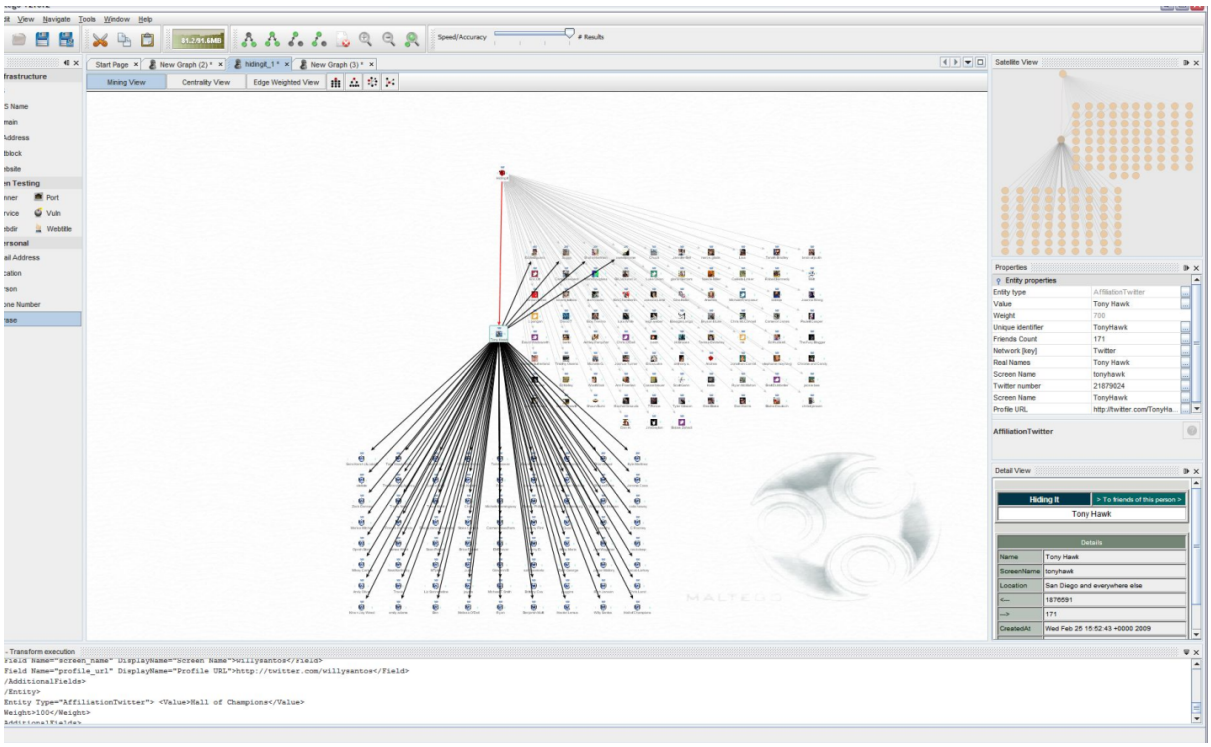

Figure 5.10 People referenced in @Tonyhawks tweets

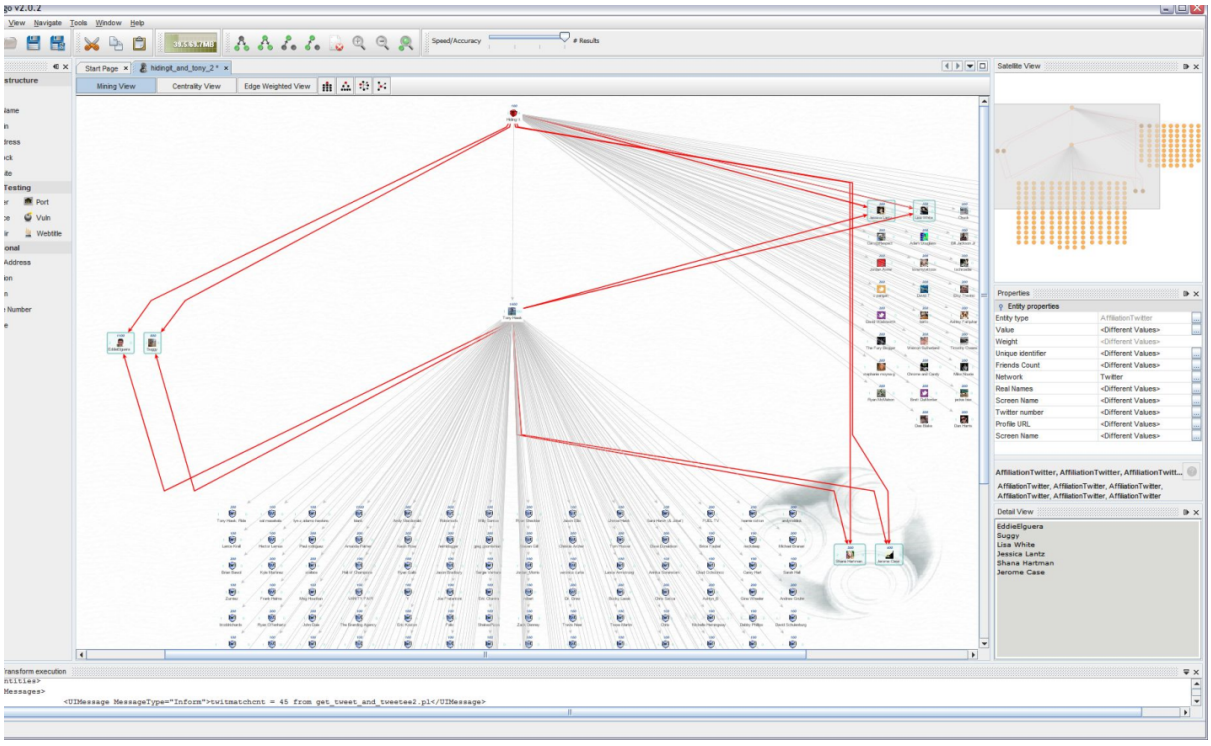

Figure 5.11 Emerging links between people Tony mentioned in tweets and follower of @HidingIt

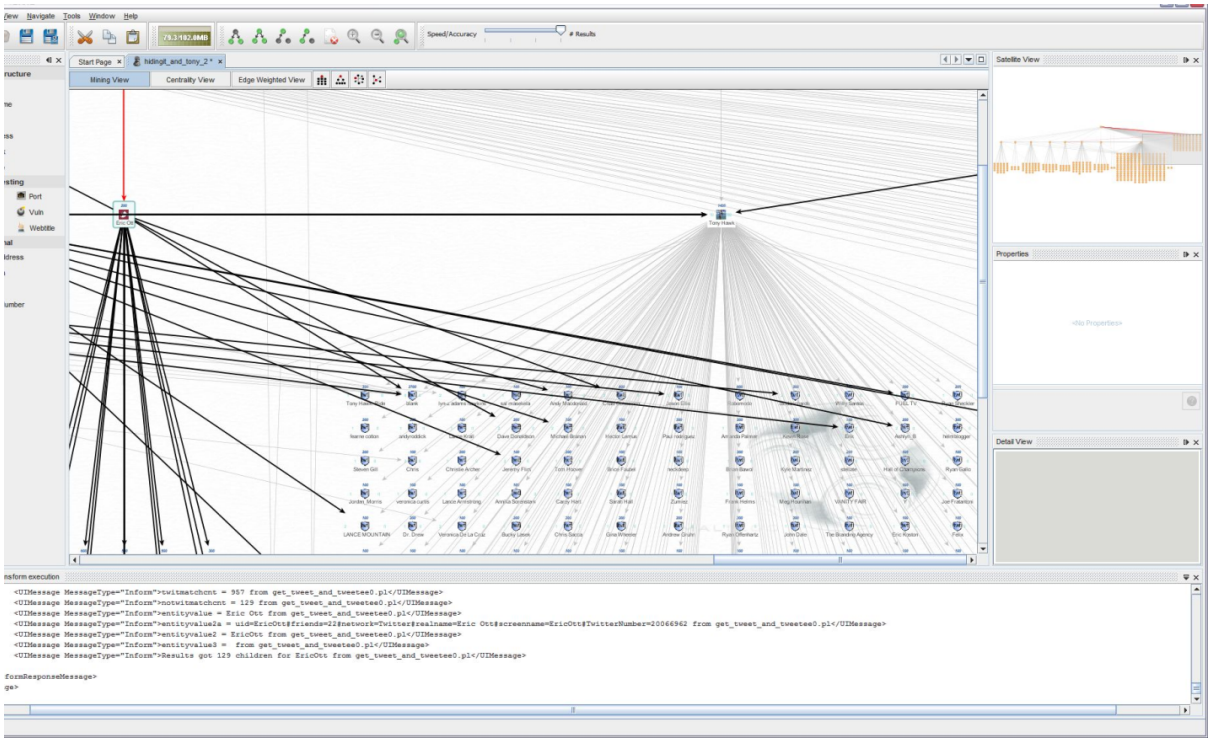

Figure 5.12 More links between people Tony mentioned in tweets and follower of @HidingIt

9. Repeat the step to generate a graph of people Tony mentioned in his tweets AND all the people they referenced in their tweets

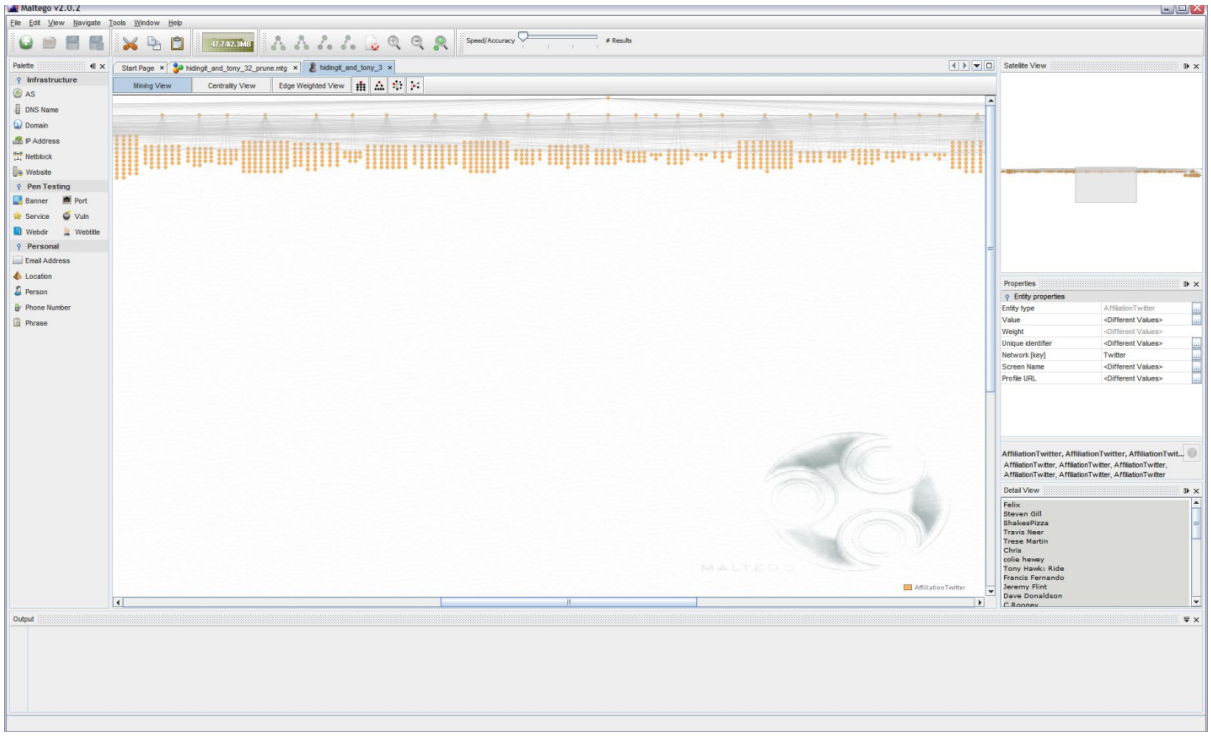

Figure 5.13 Resultant graph of people Tony mentioned in his tweets AND all the people they referenced in their tweets.

5.6 Visual Graph Exploration

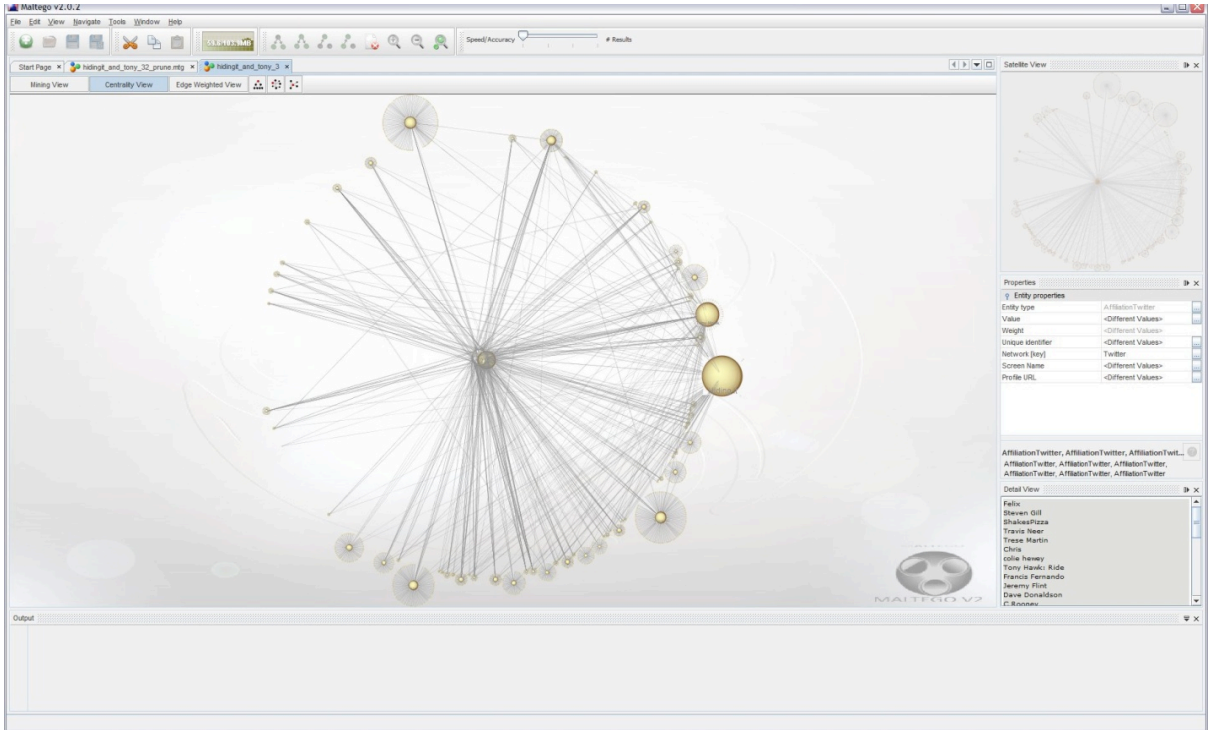

Figure 5.14 Centrality view of resultant graph of people Tony mentioned in his tweets AND all the people they referenced in their tweets.

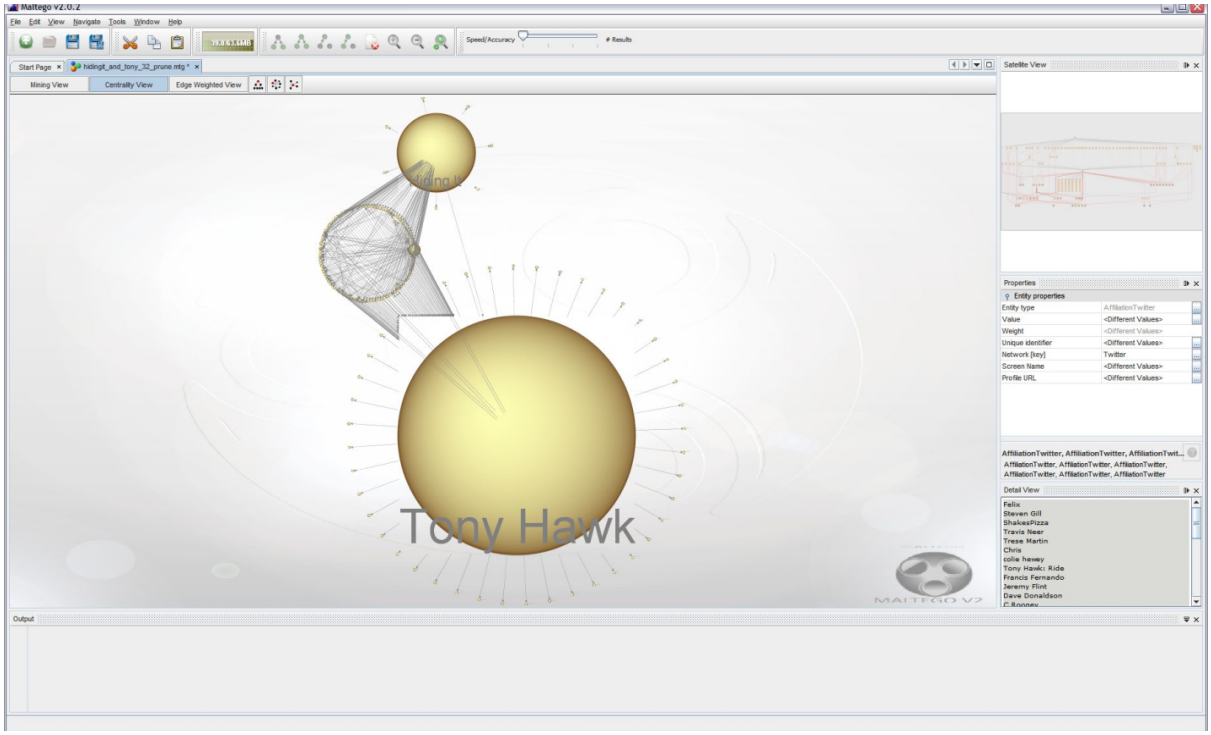

Figure 5.15 Edge-weighted resultant graph of people Tony mentioned in his tweets AND all the people they referenced in their tweets.

10. Prune the graph, removing nodes *(people)* who are not interconnected.

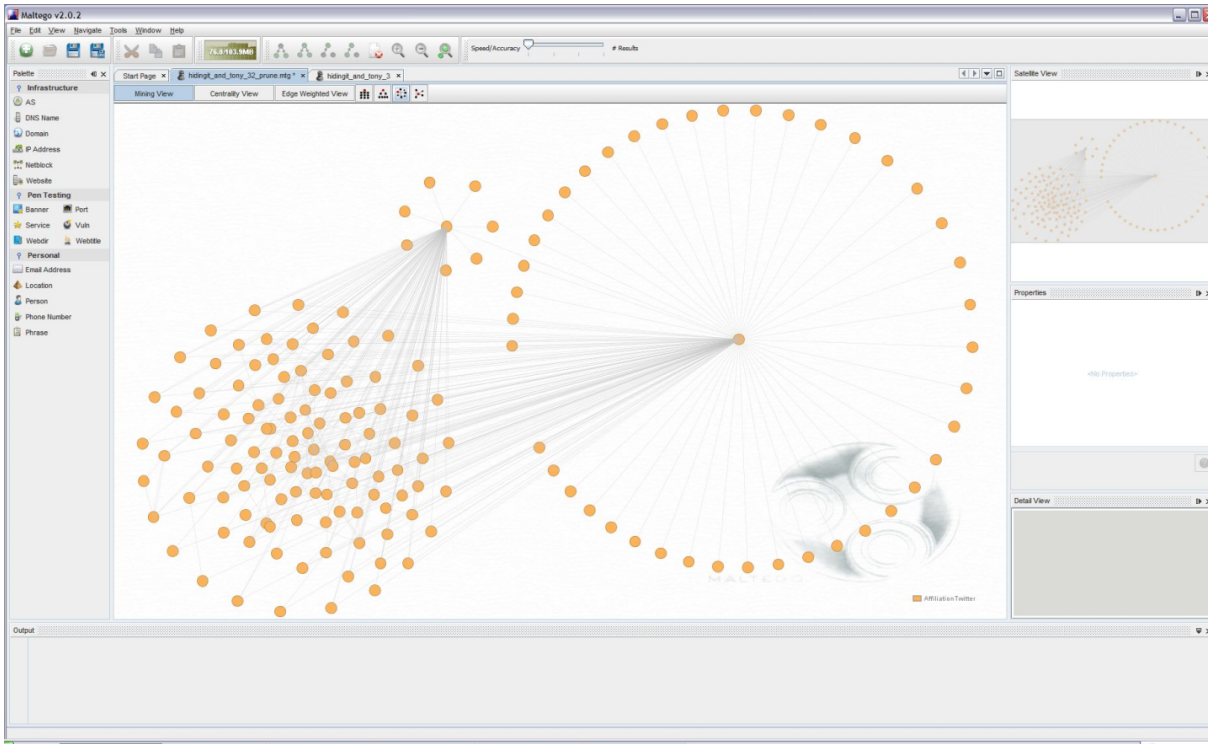

Figure 5.16 Pruned version of tree, showing connections between people referenced in Tony's tweets and friends of @hidingit

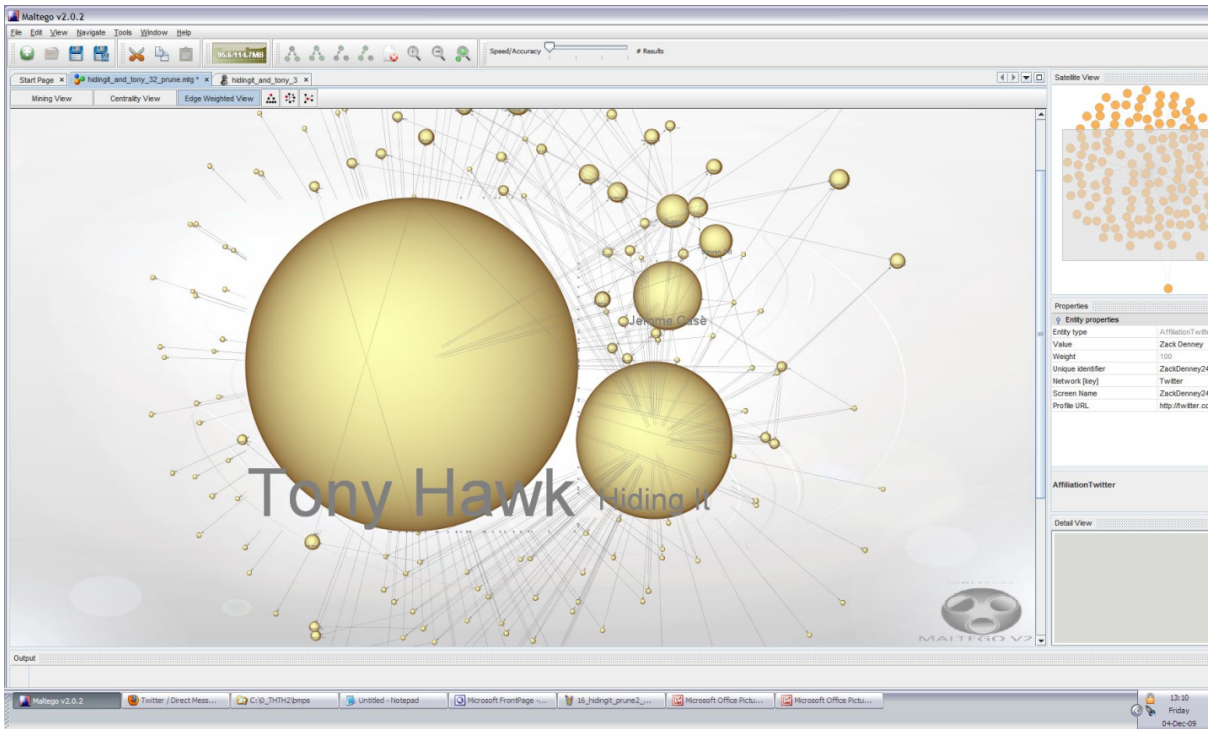

Figure 5.17 Edge Weighted version of pruned tree, showing connections between people referenced in Tony's tweets and friends of @hidingit

5.5 What does this tell us?

So what does this actually tell me. Well, it tells me that even if I hadn't been following the #THTH event, I can see that the following were pretty active talkers/talkee's about #THTH in comparison to others

- @tonyhawk (obviously),
- @SweetJerome (Tony's helper and all-round generally awesome dude),
- @Steven_Gill (read his story here) and
- @TheSuggmeister (yours truly)

If you (with no knowledge of the event) had determined this, you'd probably have drilled down to my blog and read the articles that myself and @Steven_Gill wrote. You'd have also been able to follow the links on my blog and read other hider/finder stories. You'd also have figured out that @SweetJerome pretty much ran the event for @tonyhawk. Just with these 4 pieces of information, you'd probably know everything about the #THTH event you could ever possibly want to know.

6. Case Study # 2

This case study varies from the previous as it focuses on Facebook rather than Twitter. Generally speaking, "tweets" are public, where as many Facebook users limit access to interesting details to "friends". Therefore, it is highly likely that the first task must be to win the confidence of the people you wish to enumerate.

6.1 Special Note - Facebook Terms of Services

At the date of publishing this paper (2nd August 2010), Facebook Terms of Services were clear that collecting users data first requires their consent. You are strongly advised to familiarize yourself with the terms of service as Facebook take breached of Terms Of Service very seriously.

6.2 Background

In 2010 a series of scams associated with one email address and a valid postal address in Europe attracted the attention of a local police force. Together with the local police force, we used visual data analysis as part of a tool kit in gathering intelligence based on publicly available information.

6.3 Information we started with

- Email address of scammer (Bob)
- IP Addresses of the scammer
- Name & address of the recipient of the stolen goods (Alice)

6.4 Determine location of scammer

A whois lookup put the addresses in Lagos, Nigeria. (41.220.....)

6.5 For both recipient & scammer determine if the person exists on social network sites.

Facebook transforms, such as those written as a proof of concept by Dominic White, would quickly be able to generate 3 possible results for the recipient of the goods.

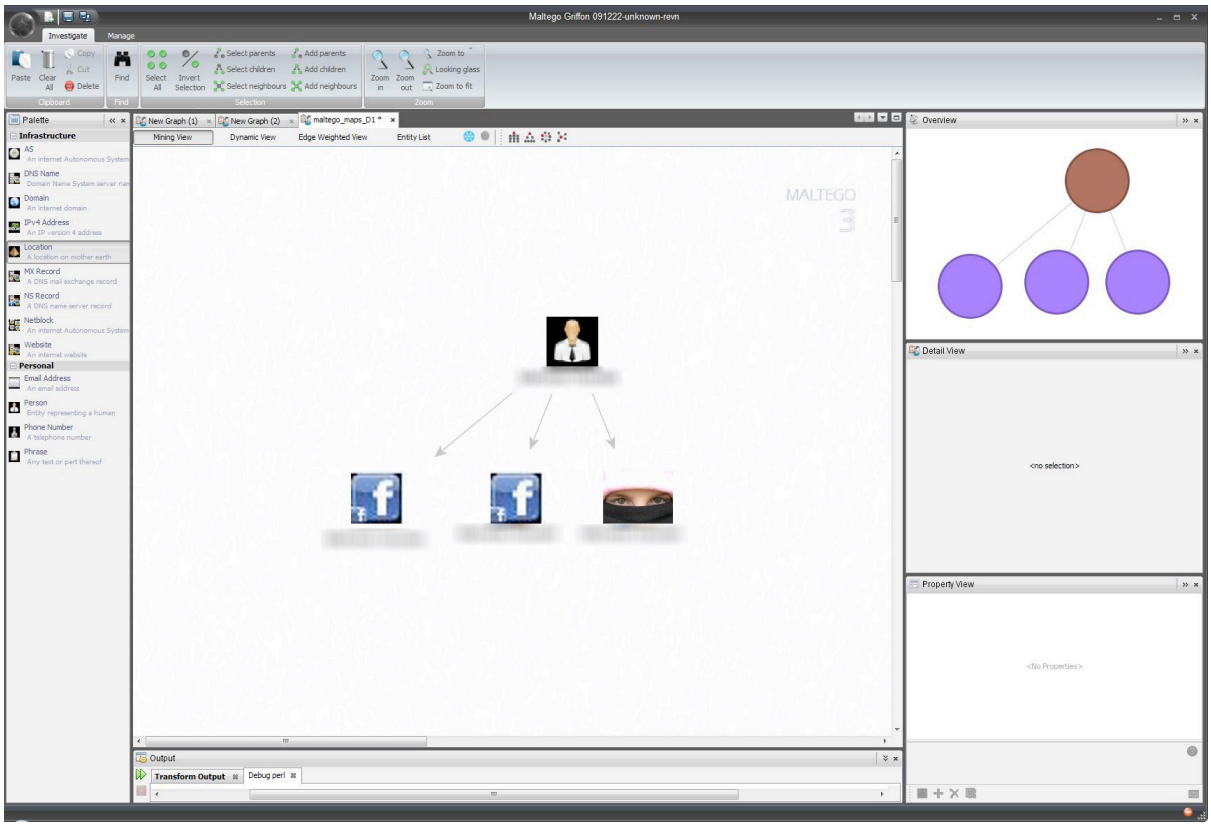

Figure 6.1 Example Person to Facebook transform

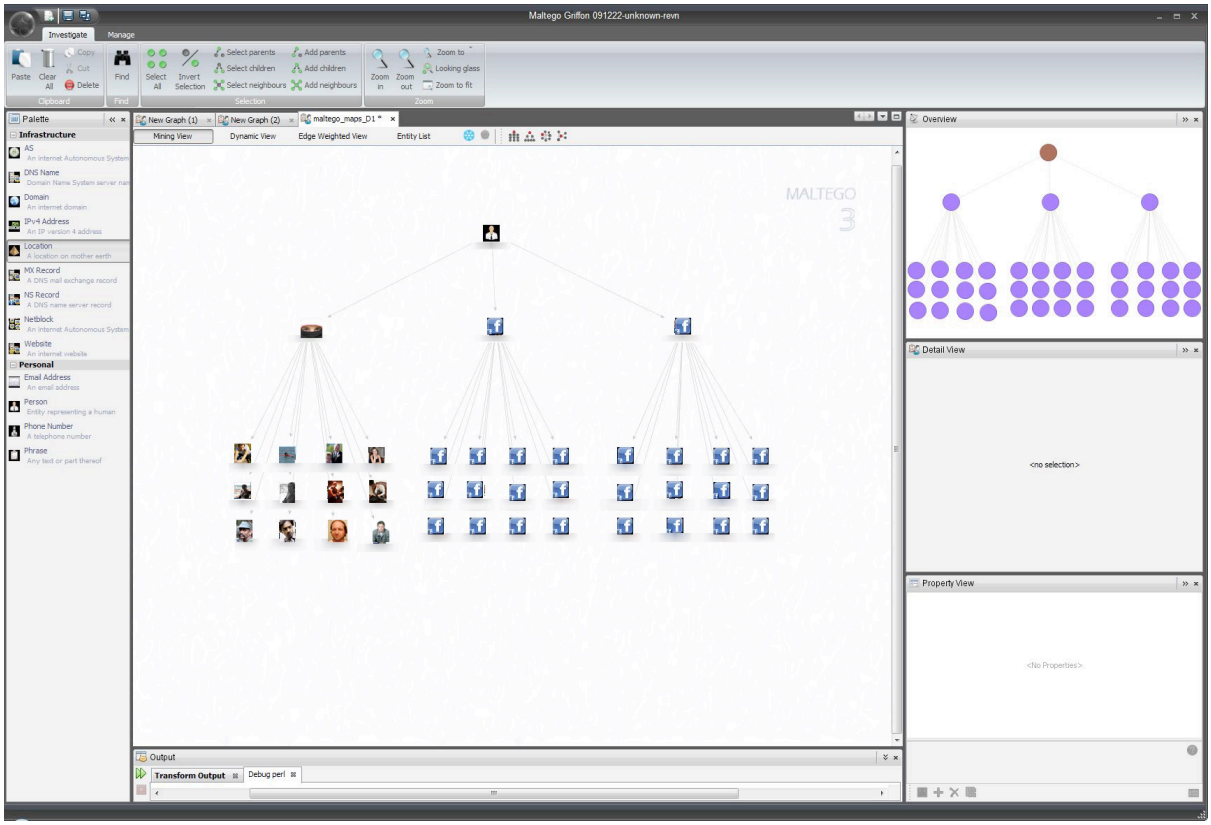

Figure 6.2 Facebook to Friends transform (note, only showing first 12 results)

6.6 Examine location of people

Another local transform to extract profile location could quickly narrow this search down to one.

Alice, based in “Newcastle” has a significant number of Facebook friends in Nigeria.

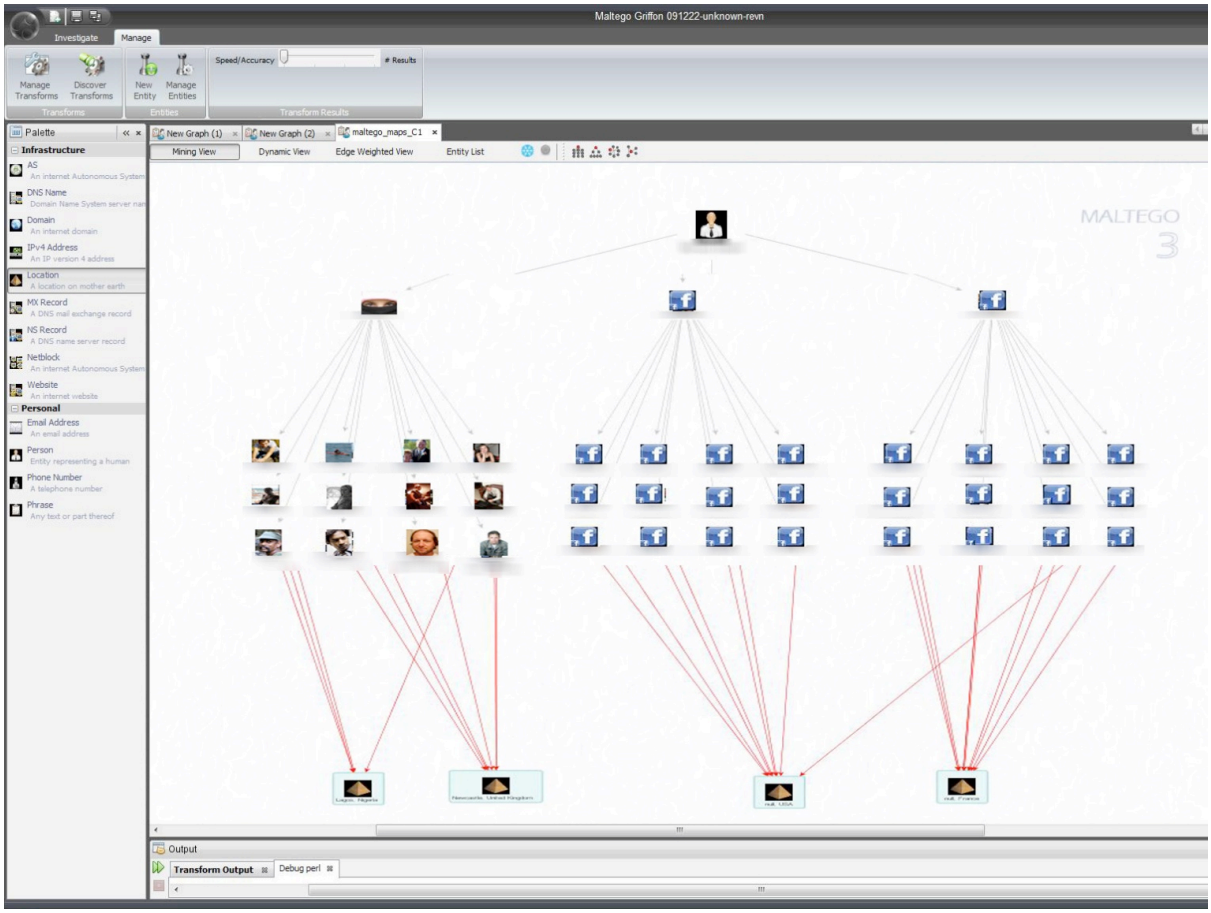

Figure 6.3 - Example Facebook to Location local transform

6.7 Drill down in details

It's possible to narrow the result set down further based on information within wall-posts and photo's for each Nigerian "friend". We performed this outside of Maltego.

This stage alone exposed some fascinating results. Photo's which seemed to feature hoards of stolen goods and comments supporting these assumptions.

6.8 But is one of these guys the scammer?

This was difficult to ascertain as the scammers email address wasn't associated with his Facebook profile.

A variety of techniques could be employed to expose any links, but the most expedient is to socially engineer the scammer to post a Facebook update.

6.9 Result

Within a short amount of time it was possible to link the scammer to the recipient of the goods and expose social relationships. From here it was possible to broaden the search identifying more and more people actively involved in scams.

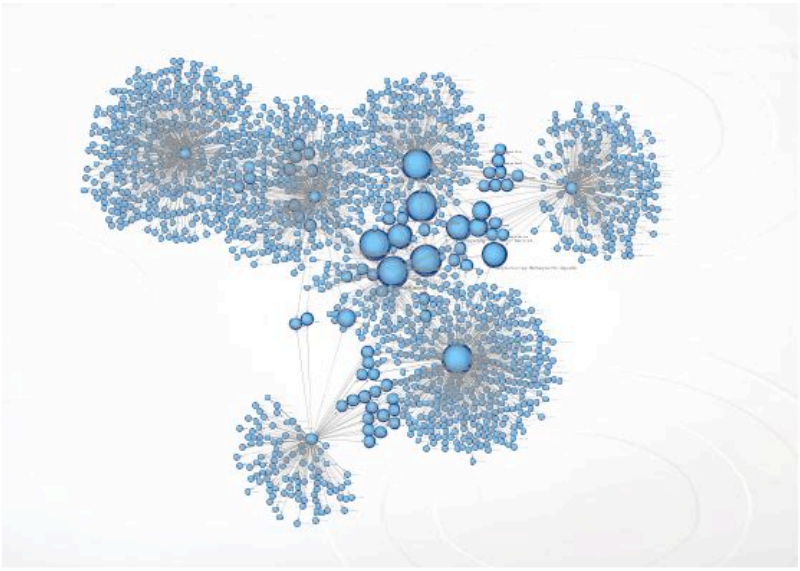

Figure 6.4 Resultant crime network and friends. In this graph, the bigger dots are more “interesting”

7. Conclusion

Data can certainly be analyzed non-visually, but visual data analysis can be effectively employed to point a researcher in the right direction.

Visual data analysis should not be viewed as a substitute for other methods of data mining and analysis, but as a complimentary practice.

Access to social networks data through API's and extendable visualization tools such as Maltego means that it is already possible for organizations and individuals to generate and analyze complex graphs with relative ease and at relatively low cost.

With little knowledge of the data set, a researcher can quickly identify key actors in a social network graph.

With more knowledge of a data set, a researcher can use interactive data visualization and other complementary techniques to uncover obscure, yet significant relationships.

As interfaces to data become more ubiquitous, individuals and organizations, both good and bad will be able to mine social data with ever greater ease.

8. Future Development

This paper has aimed to present possibilities. Individuals, organizations and agencies wishing to further explore data visualization in the context of social networks should give thought to;

- Collaboration with social network sites to ensure operation within an agreed Terms of Service.
- Creation of a full set of social network transforms.
- Creation of transforms to explore photo sharing site.
- Linking to private data sources (e.g. corporate databases, police records).

References

[1] Scott, John P. 2000. Social Network Analysis: A Handbook. London: Sage Publications Ltd,

[2] Scott, John P. 2000. Social Network Analysis: A Handbook. London: Sage Publications Ltd,

[3] http://www.wikipedia.org/wiki/Sociogram

[4] http://www.wikipedia.org/wiki/Sociogram

[5] http://hci.stanford.edu/jheer/projects/vizster/

[6] http://www.analytictech.com/ucinet/help.htm

[7] http://www.pcmag.com/article2/0,2817,2361820,00.asp

[8] http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf

[9] http://www.clickymedia.co.uk/2009/12/facebook-reaches-350-million-users/

[10] http://jasonvanorden.com/twitter-opportunity

[11] http://www.myspace.com/pressroom?url=/fact+sheet/

[12] http://paidcontent.co.uk/article/419-bebo-sold-to-criterion-armstrongs-staff-memo/

[13] http://pres.linkedin.com/about

[14] http://www.friendster.com/info/index.php

[15] How different are young adults from older adults when it comes to information privacy (April 14 2010)

[16] Gross, R., Acquisti, A. "Information Revelation and Privacy in Online Social Networks (The Facebook Case). Carnegie Mellon University, 2005.

[17] Govani, D., Pashley, H. "Student Awareness of the Privacy Implications When Using Facebook". Carnegie Mellon University, September 2007.

[18] Solove, D. ""I've Got Nothing to Hide" and Other Privacy Misunderstandings of Privacy", George Washington University, 2007

[19] http://query.nytimes.com/gst/fullpage.html?res=9E0CE3DD1F3FF93AA3575BC0A9609C8B63

[20] http://gregsadetsky.com/aol-data/

[21] http://www.readwriteweb.com/archives/news_patterns_finding_hidden_threads_in_everyday_n.php?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+readwriteweb+%28ReadWriteWeb%29

[22] Fleming, N. D; (1995), I'm different; not dumb Modes of presentation (V.A.R.K.) in the tertiary classroom, in Zelmer, A., (ed.) Research and Development in Higher Education, Proceedings of the 1995 Annual Conference of the High Education and Research Development Society of Australasia (HERDSA), HERDSA, Volume 18, pp 308 - 313.

[23] http://www.wikipedia.org/wiki/Visual_thinking

[24] Fry, Ben. Visualizing Data. Sebastopol, CA: O'Reilly Media, 2008.

[25] http://en.wikipedia.org/wiki/Named_entity_recognition

[26] Tucker, A, B; (2004), Computer Science Handbook, Second Edtion. Chapman & Hall/CRC, pp20 - 22.

[27] http://www.paterva.com/web5/

[28] http://www.paterva.com/web5/documentation/localtransforms.php

SecurityG33k

www.securityg33k.com | TheSuggmeister@gmail.com | twitter.com/TheSuggmeister