Anonymous Wikipedia edits from the US Senate
Milad Gholami (mgholami@cs.umd.edu)

Introduction:
A while ago there was a news about someone with an IP address belongs to the US Senate edited the Wikipedia page of Edward Snowden [1]. We can see here [2] the IP address range belong to the US Senate. These anonymous edits are usually based on personal opinion of people who maybe work there. Generally, the data of these anonymous edits are not hard to extract and might have interesting information.

Dataset:
We know that if someone makes an edit to a wiki page anonymously (without logging in), it will be stored by IP address. And you can see edits made by an IP address in pages like [3]. To create this dataset I crawled (with a bash script using curl) Wikipedia edit history pages made by each IP address from the US Senate IP address range (156.33.0.0 - 156.33.255.255) and then I extracted necessary information by python scripts. And imported the result to NodeXl.

In my graph the nodes are Wikipedia pages edited anonymously by an IP address from the US Senate IP address range (156.33.0.0 - 156.33.255.255) and they are connected if they edited by the same IP address. It has around 1400 nodes and around 5600 edges between them. It is an undirected graph. This is how the data originally looks like:

![Initial view of the data](http://nodelink.com)

Figure 1 Initial view of the data

Headline1: Distinguish wiki pages correspond to the same area
I used Girvan-Newman with clustering algorithm with Fruchterman-Reingold layout and after some steps I got to this view of the data:

![Graph Clusters](http://nodexl.codeplex.com)

**Figure 2 Graph Clusters**

As one could expect some of these clusters are clearly pointing to a specific area or state. Like I examined the light blue one in the middle and examples of pages found there are: [4], [5], [6], and [7]. These all have a connection to the state of Maine. They are either a location or a political person from Maine. And those edits has been made by different people in the Senate and seems like they all share interests about the state of Maine. This case is also true about some other clusters like the light orange one on the top right are all related to Maryland.

**Headline 2: Some clusters correspond to personal interests and are not shared with others**

I used the same clustering algorithm and layout but this time I wanted to see each cluster individually.

![Closer look to clusters](http://nodexl.codeplex.com)

**Figure 3 Closer look to clusters**

For instance the cluster in the third column from the left are all about Bollywood movies [8]. The top on the 4th column from the left is the pages created by one IP address (it is a clique) and you can find things
like Foundation (novel) or Universal Soldier (1992 film). The second one on the same column is again a clique and you can find things like Film noir and Near-Earth asteroid. These are topics that are edited by only one person and has no relation with pages other people edited.

**Headline 3: “United States Senate Committee on Finance” Wikipedia page has the most anonymous editors from the US Senate**

I calculated the degree of each vertex and considered it as the size of vertex. Then I sorted the size column and this is the first one which is not from a clique.

![Image](http://nodexl.codeplex.com)

*Figure 4 the most edited wiki page*

**NodeXL Critiques:**

I tried to use both Gephi and NodeXL for this Homework. Clearly with NodeXL it is easier to import the data, specify the size, color, shape, thickness, and opacity of the edges and vertices. And the clustering algorithm seems to work better. Also there are lots of recourses to get help while working with NodeXL. One another positive point is that it is inside excel which makes it easier to work on the data. Also, I could easily copy my data from other excel sheets directly to NodeXL edges or vertices.

However, I faced several issues while doing this homework. One clear disadvantage of NodeXL is that, it does not have a mac version. Which made it so hard for me and I had to borrow a windows laptop. Also, it needs a lot of patient to work with NodeXL because it crashes a lot. There were also issues about importing social network data to NodeXL. Whenever I tried to download a twitter data it was reaching the limit so fast and I had to pause. The other thing I wanted to use was the facebook fan page data. It has options to add edge for every two person that like a post on the page. Clearly it is not a good idea and if a post has 100 likes, you will have $2^{100}$ edges. I wanted to make it to add an edge between two people if they like at least two posts or more from this page which I couldn’t. NodeXL could be much more helpful by improving the facebook import capabilities and adding more options to it. Then I tried to get my gmail data. At first it made me to set up my gmail account on my friends laptop. Then it was really slowly importing the email data and was almost no useful. In this regard Gephi was much more
helpful. It was much faster in downloading the emails, and I didn’t have to import my Gmail to Outlook first. And it didn’t crashed even though I had a lot of emails. Generally, NodeXL is much slower than Gephi. Maybe because Gephi is not limited to Microsoft Excel.

References:

2- http://wolfsbane.toolserver.org/~overlordq/cgi-bin/whois.cgi?lookup=156.33.241.5
4- http://en.wikipedia.org/wiki/Chewonki_Foundation