Using Social Interactions Network Graph and Centrality to Identify KeyPlayers

Alex Pongpech[a]

[a] Big Data Engineering Graduate Program, Dhurakijpundit University, Thailand

E-mail address: alex.pon@dpu.ac.th

Abstract

Topic on Trend has been made more popular recently with the published Food trend 2016 by Google. Prior to the social network era, difficulty in predicting and identifying trend are difficult at best. This is mainly due to difficulty of gathering data from the public to do the analysis. Given that graph can be utilized to modeled social network users and their relationships, and that graph algorithms are very mature. The possibility of utilizing graph algorithms to analyze social network users to help identifying trendsetters is worth investigating. In this paper, the aim is to apply graph theory to model interactions on social network. The model can then be utilized to identify keyplayers based on the Betweenness centrality and PageRank centrality. Finally, based on PageRank algorithm, vertexes ranking is implemented using python.

Keywords: Trend, Social Network, Graph, Centrality, Keyplayer, PageRank

1. Background/ Objectives and Goals

The term "trend" implies a general direction in which something is developing or changing. Understanding trend and the ability to predicting trend have always been the goal of every business. When discussed in the economy context, the terms such as bear market and bull market have been used to describe the market trend. When discussed in fashion industry context, trend implies the direction which people are dressing (Kieselmann, 2014). In any given context, ability to predict and identify these trends has always been sought after.

Knowing trend provides business with better information to manage demand and supply. If a particular product is on the rising tide of the trend, then the company might benefit best if it has arranged to buy/stock raw materials in advance. On the other hands, if it is known that the tide of the trend is on the wane then company should not be stocking raw material and should start planning on new product. From financial market, fashion design, entertainment industry to technology industry, each has more or less benefit from knowing trend in advance. Some have failed to capitalized from following the trend, and some have actually failed completely by following the wrong trend. Some of example questions that can be answered through to estimate trend are as follows.

- demand and supply (given the climate, is this trend going to cost more?)
- health related (the money spending on health issue, the growth of kids)

- financial (the amount saving or spending in the food industry)

For food industries, knowing trend in advance is even more important. The ability to identify and predict trends are highly important in Food Industry. Raw material use in food industry is usually have a shelf life. Not only that, the price of such commodity is fluctuated. For example, agricultural productivity is heavily based on climate, and will be price accordingly. Knowing food trend in advance is also benefit healthcare industries. Rising sugar consumption can indicate possible increase in health medication in the near future. These trends are not only providing competitive edges but can be utilized to facilitate food sustainability such that food can be produced, processed, bought, sold and eaten in ways that provide social benefits, globally and locally.

Under normal circumstance, the difficulty in predicting and identifying trends are difficult at best. Apart from complex statistical techniques, it is due to the difficulty of gathering data from public to do the analysis. When examining patterns or trends over time, several sources of data such as food balance sheets (FBSs), household budget surveys, or individual dietary surveys (IDSs) can be used. Unfortunately, obtaining such data is a complex task. Not only, an appropriate amount of data is needed, the data must be collected over a substantial period of time before trend can be estimated. Such delay and complexity in collecting data could have a big impact on industry trying to identify new trends in a timely manner.

Given a large amount and various types of data that can be obtained much quicker and broader than normal surveys through the web, the question is can we examine patterns and identify trends using data scarping from the social web instead of the traditional surveys? Google(Google, 2016) has illustrated that through analyzed google web queries, it is possible to identify various food trends. Through such data, Google were able to find food trend characteristics with various associate factors such as geography, seasonal, top associated keywords, and top day for searching. Google has demonstrated convincingly that data from the web can be utilized effectively to help examine and analyze trends.

The amount of data that exploded into the web can be traced to the rise of social network web such as web-blogs, Facebook, Tweeter, and Instagram. These social network web applications allowed users to uploaded and downloaded a huge amount of images, text, audio, and video onto the web. Given these large amount and various types of data that can be extracted from the social web, can we identify data types and their potential usage to facilitate trend analysis? One manner is to investigate relationships and possible structures that along with data that have been uploaded and downloaded by the users.

The Social network analysis (SNA) (Grandjean, 2016) is the process of investigating social structures through the use of network and graph theories. It characterizes networked structures in terms of nodes (individual actors, people, or things within the network) and the ties, edges, or links (relationships or interactions) that connect them. Examples of social structures commonly visualized through social network analysis include social media networks, friendship and acquaintance networks, collaboration graphs, kinship, disease transmission, and sexual relationships. These networks are often visualized through sociograms in which nodes are represented as points and ties are represented as lines.

The question is then can we utilized these types of social web data to identify trends and examine patterns? Given that graph can be utilized to modeled social network users and their relationships, and that graph algorithms are very mature. The possibility of utilizing graph algorithms to analyze social network users to help identifying and possibly predicting trend is worth investigated. One manner to achieve this is through keyplayers or trendsetter on the social network. These trendsetters are usually moving in the direction that parallel with the rising trend. For example, if we can identify food pages or culinary persons that are the trendsetters then we might have rewarding targets to mine social interactions network data. In this paper, the aim is to present techniques that have been utilized in identifying keyplayers. A social interactions network model is also given. Finally, a vertexranking algorithm based on PageRank is presented as a potential useful technique to investigate keyplayers through centrality of the social network.

2. Methods

From the last section, the potential of utilizing data from social web to facilitate trend estimation has been discussed. In this section, a more detail of social network analysis using graph mining is given. Data that associated with social interactions is discussed and methodology to utilized such data with graph structure is proposed.

2.1 SING (Social Interactions Network Graph)

Mining social networks has a long history in social sciences. Zachary's PhD (Zachary, 1977) observes social ties and rivalries in a university karate club, where he found that conflicts within the group led the group to split. This split can be explained by a minimum cut in the social network, and it can be easily detected through social structure on social network graph as illustrated in Figure 1.

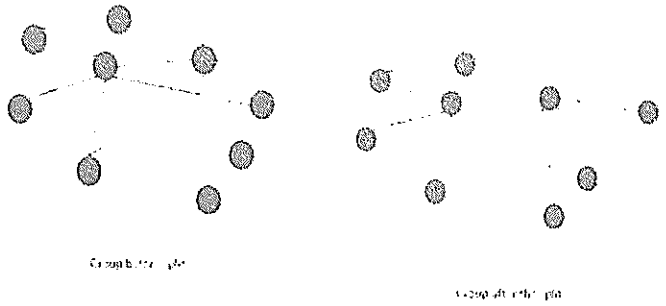

Fig 1: Graph structure

Social network is most effective modeled using graph theory, a brief discussion on graph theory (Tutte, 2001) is given as follows. The fundamental concept of graph theory is the graph, which is a mathematical object. A graph can be denoted as G(V, E) or G = (V, E) which consists of set of vertexes V together with a set of edges E. Vertexes are also known as nodes, points and (in social networks) as actors, agents or players. Edges are also known as lines and (in social networks) as ties or links. An edge e = (u,v) is defined by the unordered pair of vertexes that serve as its end points. Two vertexes u and v are adjacent if there exists an edge (u,v) that connects them. An edge e = (u,u) that links a vertex to itself is known as a self-loop or reflexive tie. The number of vertexes in a graph is usually denoted n while the number of edges is usually denoted m. As an example, the graph depicted in Figure 2 has vertex set V={a,b,c,d,e.f} and edge set E = {(a,b),(b,c),(c,d),(c,e),(d,e),(e,f)}.

Fig 2: Graph representations

When used to represent social networks, each line is used to represent instances of the same social relation, so that if (a,b) indicates a friendship between the person located at node a and the person located at node b, then (d,e) indicates a friendship between d and e.

A graph can be analyzed in partial where each part is called subgraph. A subgraph is denoted as G'=(V', E') of G = (V, E) where $v' \subset V$ and such that $(v_1, v_2) \in E'$ if and only if $(v_1, v_2) \in E$ and $v_1, v_2 \in V'$. Figure 1 also represent a graph that can be partition into two subgraphs.

Given a connected graph, we give more graph terminology as follows. A path is given as a sequence of vertexes $v_1, v_2, \ldots, v_n$ that there is an edge from each vertex to the next vertex in the

sequence. The length of the path is n-1, which is the number of edges along the path. Two vertexes v_i and v_j are defined as connected if and only if there is a path that starts with v_i and ends with v_j.

Information that encoded in graph can be extracted through characterizing graph structure. The number of edges coming out or going in to each vertex is call degree. The higher the degree of a given vertex implies a large connectivity with other vertexes. A vertex that has a higher degree in social network is more likely to have a higher influence that a lower degree vertex.

There are times that finding out how fast the information flow through social network graph can be useful. Information usually flows through social network quicker through higher degree vertexes than lower ones. The distance between two high degree vertexes can also tell us how easy information can propagate through social network graph. The shortest path between two high degree vertexes give us mean to compare how fast information might be propagated through a given social network graph. For example, two high degree vertexes with a shorter path could facilitate information propagated through social network quicker than the other two similar degree vertexes that has a longer shortest path. The shortest path length between two vertexes *i* and *j* is the number of edges comprising the shortest path (or a shortest path) between *i* and *j*.

Graph can also be easily represented as matrix where each graph has associated with it an adjacency matrix. That can be represented as a binary nxn matrix A in which $a_{ij} = 1$ and $a_{ji} = 1$ if vertex v_i is adjacent to vertex v_j, and $a_{ij} = 0$ and $a_{ji} = 0$ if there are no relations between the two vertexes. This relation can be represented using table as illustrated in Figure 2.

In social network community such as Facebook, information that can benefit trend analysis such as friendship and informal contacts among friends, collaboration and influence in companies, organizations, communities are described through connection between them. If two or more domains have social network relations, they have edges connected between them. These sources of rich social network structure are emerging at a very rapid pace as more people participating on the internet. Content creation, blogging, social media, and electronic markets are very active and full of information. Formal channels such as Facebook, LinkedIn, and Line have seen explosion in users' participations. These arbitrary relations between people or various elements on the social network can be captured using graph as illustrated in Table 1.

Table 1: Data instance to Graph instance

Data Instance		Graph Instance
Element	⟹	Vertex
Element's Attributes	⟹	Vertex Label
Relation between elements	⟹	Edge, directed and undirected
Type of relation	⟹	Edge Label
Relation between a set of elements	⟹	Hyper Edge

A labeled graph, G = (V, E), is a finite series of graph vertices V with a set of graph edges E of 2-subsets of V. Given a graph vertex set $V_n = \{1, 2, ..., n\}$, the number of vertex-labeled graphs is given by $2(n(n-1)/2)$. A labeled graph can use numeric number instead of just text label. Such graph is called weighted graph, which is defined as a graph for which each edge has an associated weight, usually given by a weight function w: E→R

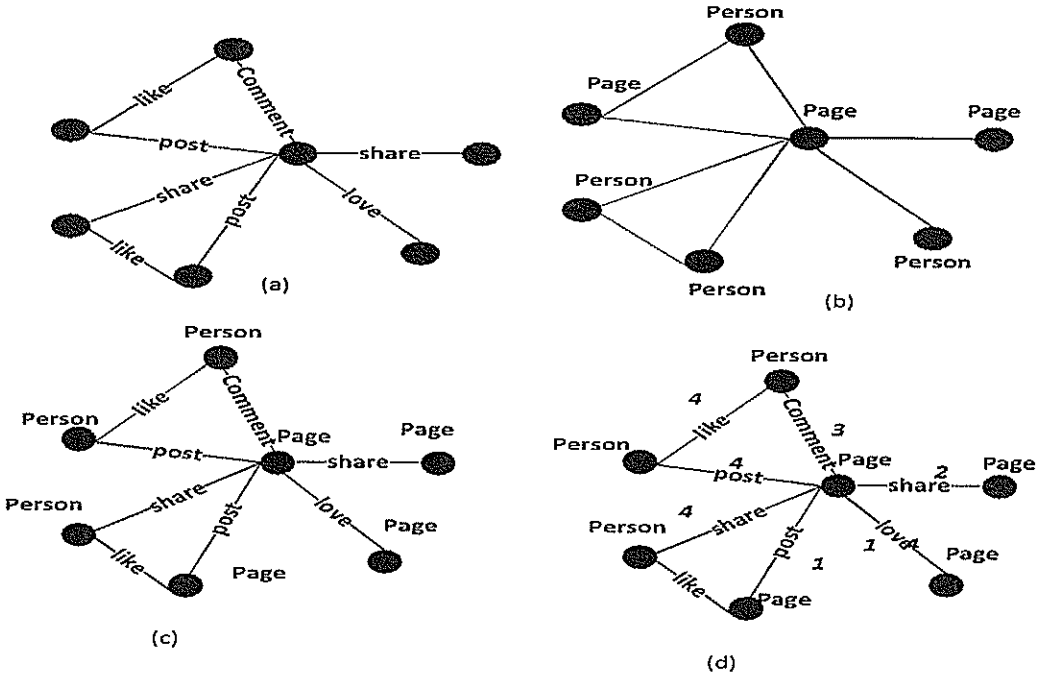

Fig 3: Social Network Graph

Figure 3 illustrated how data instances in Facebook can be represented as a graph. Starting with a social graph with labeled edge, a social graph with labeled vertex, a social graph with both vertex and edge labeled, and a social graph with weighed and labeled edge with labeled vertex. From the illustrated graph, we can see that using undirected graph model is not sufficiently enough to express direction of activities on the social network.

Using a directed graph can enable us to capture actors more adequately. Directed graph (digraph) is defined as G= (V, E) consists of a non-empty set, V, of vertexes, and a set $E \subseteq V \times V$ of directed edges. Each directed edge $(u, v) \in E$ has a start vertex u, and an end vertex v.

Interactions between two vertexes in a social network graph are usually not limited to simplex. A one direction, digraph, is then not adequate for capturing semantic rich information of Social interactions in the network. A multigraph allows two or more directed edges to be specified and thus enable us to express and capture duplex interactions between vertexes. A multigraph, G = (V, E, f), is defined as a set of vertexes, V, a set of edges, E, and a function f: E→ $\{\{u, v\} \mid u, v \in V \text{ and } u \neq v\}$.

2.2 Centrality

In social network analysis, it is very important to be able to identify key or central entities in the social network. There are a number of measurement that can be applied to measure key or central vertexes from the social network graph. The term centrality is a measure of how many connections one vertexes has to the others. In this paper, we give four type of centrality measurements as follows.

First, degree centrality refers to the number of ties a vertex has to the other vertexes. For a digraph, the degree centrality is composed of in-degree centrality and out-degree centrality. Those vertexes with higher in-degree centrality can be categorized as prominent vertexes, and those with higher out-degree can be categorized as influential vertexes. Second, betweenness centrality is an indicator of a vertex's centrality in a given social network. It can be compute by calculate the number of shortest paths from all vertices to all others that pass through that node. The betweenness centrality enable us to determine vertex that is the most direct route between the other two vertexes. Third, closeness centrality is the mean length of all shortest paths from a vertex to all other vertexes in the social network. A person on a social network with a high closeness centrality can spread information to the others on the social network quicker. Fourth, instead of using the eigenvector centrality, we choose to use PageRank centrality instead as it handles digraph more effectively. In a directed acyclic graph, eigenvector centrality becomes zero, even though the vertex can have many edges connected to it (Zafarani, Abbasi, & Liu, 2014). A PageRank measurement can be used instead to determine how well a person connected to other well connected person. In this paper, we are focusing only on the betweenness centrality and the PageRank centrality.

Betweenness centrality is a measurement concept use for analysis of social networks. For undirected graph, Freeman (Freeman, 1980) proposed to derived betweenness centrality for undirected graph from the column totals of a single matrix of the numbers of pairwise dependencies of each point on every to the point as a platform to reach third points. He also shows that for undirected graphs, the sum down the columns of the pair-dependency matrix D is a measure of betweenness centrality of the points given as

$$\sum_{i=1}^{n} d_{ij}^{*} = 2C_B(v_j) \quad (1)$$

Where $C_B(v_j)$ is the betweenness of point j, and d_{ij}^* is the pair dependency v_i on v_j. If D is defined as a matrix, $D = (d_{ij}^*)$, then we can extract important information about gatekeeper with respect to each other point from the matrix. A gatekeeper is an important member of the social network community who has either formal or informal influence with the culture (Foster, Borgatti, & Jones, 2011). It is also possible that a gatekeeper can be think of as a trend starter. In order to determine the betweenness centrality for digraph, White (D.R. White, 1994) extends Foster's measure of betweenness centrality to cover digraph by taking the number of points with outgoing edges, the number with incoming edges, and the number of points with reciprocated edges into consideration. Let n_o be the number of points with outgoing edges and n_i be the number of incoming edges, and n_s be the number of points with reciprocated edges, White (D.R. White, 1994) gives the betweenness centrality of the most centralized star for a directed graph as follows.

$$C_B(v_j) = (n_i - 1)(n_o - 1) - (n_s - 1) \quad (2)$$

PageRank is a method for computing a ranking for every web page based on the graph of the web (Page, Brin, Motwani, & Winograd, 1999). Given u be a set of web page, and let F_u be the set of pages u points to and B_u be the set of pages that point to u. Let $N_u = |F_u|$ be the number of links from u and let c be a factor to keep the total rank of all web pages constant. A definition of PageRank, $R(u)$ with adjusted rank source $E(u)$, vector over the web pages corresponds to a source of rank, is given as follows.

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v} + cE(u) \quad (3)$$

3. Results

3.1 SING

Let $V = \{v_1, v_2, \ldots, v_k\}$ be a set of entities on a given social network

Let $E = \{(v_i, v_j) | v_i, v_j \in V(G)\}$ be a set of social interactions between two or more social network entities

Let $L_v = \{l_{v1}, l_{v2}, \ldots, l_{vm}\}$ be a set of vertex labels describe entities where $L_v(V) = \{L(v_i) | \forall v_i \in V(G)\}$

Let $L_E = \{l_{e1}, l_{e2}, \ldots, i_{en}\}$ be a set of edge labels describe social interactions between entities where $L_E(E) = \{L(v_i, v_j) | \forall (v_i, v_j) \in E(G)\}$

Let L be the vertex labeling function where $L: (V \rightarrow L_v) \cup (E \rightarrow L_E)$

Let r be a function express direction of the social interaction from E(G) to the set of all unordered pairs of two elements of V(G) where $r: E \rightarrow VxV = \{(u, v) | u, v \in V\}$.

We then define a social interactions network graph, S, as an ordered quintuple (V(S), E(S), $L_v(S)$, $L_E(S)$, r). An example of a social interactions network graph is illustrated in Fig 5.

Example

Given $V = \{v_1, v_2, v_3, v_4, v_5, v_6, v_7\}$ and $E = \{ (v_1, v_6), (v_2, v_7), (v_4, v_5), (v_4, v_7), (v_5, v_4), (v_6, v_6), (v_6, v_1), (v_6, v_7), (v_7, v_1) (v_7, v_2) (v_7, v_3) (v_7, v_5) \}$

$L_v = \{$person, page$\}$ where $L_v(v_1) = L_v(v_5) = L_v(v_6) =$ person and $L_v(v_2) = L_v(v_3) = L_v(v_4) = L_v(v_7) =$ page

$L_E = \{$comment, like, post, share, love$\}$ where $L_E(E) = \{((v_1, v_6)$, like), $((v_2, v_7)$, like), $((v_4, v_5)$, post), $((v_4, v_7)$, post), $((v_5, v_4)$, like), $((v_6, v_6)$, post), $((v_6, v_1)$, comment), $((v_6, v_7)$, post), $((v_7, v_1)$, comment), $((v_7, v_2)$, share), $((v_7, v_3)$, love), $((v_7, v_5)$, share$)\}$

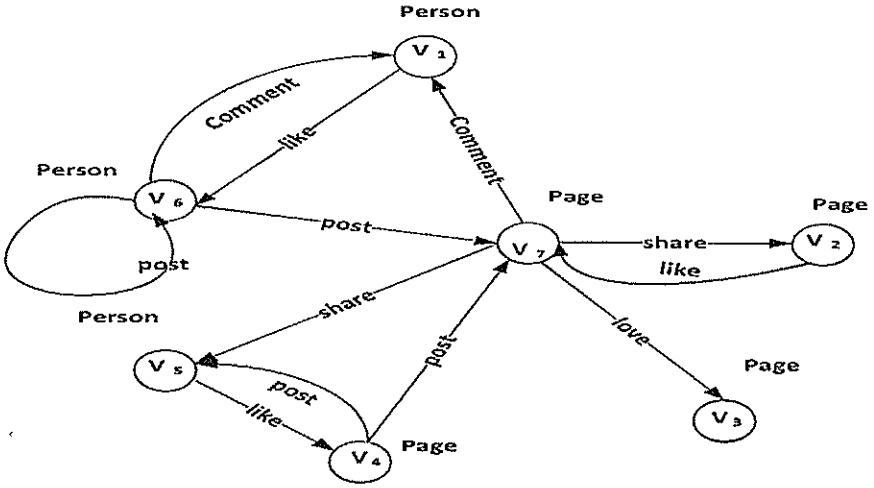

Fig 4: SING

In this part, we applied the betweenness centrality of the most centralized star for a directed graph given in Figure 5, and found that the maximal betweenness is 11. Figure 5 shown the number of incoming edges and outgoing edges.

$C_B(v_1) = 2$ $C_B(v_2) = 0$ $C_B(v_3) = 0$ $C_B(v_4) = 1$ $C_B(v_5) = 1$ $C_B(v_6) = 1$ $C_B(v_7) = 11$

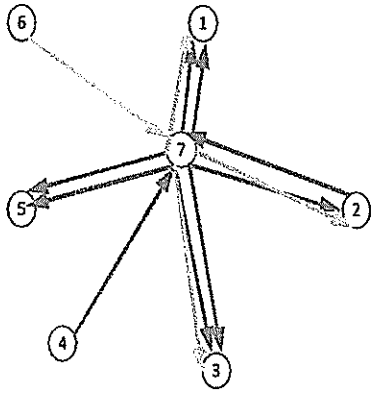

Fig 5: Maximal betweenness centrality of node v_7

Given that web pages are also considered vertexes in the web graph, we should be able to readily apply PageRank algorithm to identify power vertex in the social network graph. A vertex is *important* if there are many vertexes connected to it, and these connected vertexes must also be *important* themselves.

Let assume vertex v_7 has vertex $v_1 \ldots v_n$ which has in-degree edge connect to it. The parameter d is a damping factor which can be set between 0 and 1. In this paper, we select d to 0.85. Also $C(v_n)$ is defined as the number of outgoing edges from v_n. The VertexRank of a vertex v_7 algorithm and result is given as follows.

```
In [34]: n_vertexes = 7 # numbering vertexes v1 through v7 as 0 to 6
         M_counts = np.zeros((n_vertexes, n_vertexes)) # will hold the number of link
         counts (assumed 1 or 0)
         # columns = starting page, row = destination page, i= M_ij = whether or not t
         here is a link from j to i

         M_counts[1,2] = 1 # vertex 3 (A in the graphic) is a sink because it has no o
         utgoing links at all;
         # however, M cannot contain an all-zero column, so do as if A was linking to
         all other pages (ie put 1's everywhere)
         M_counts[5,0] = 1 # v1->v6
         M_counts[0,5] = 1 # v6->v1
         M_counts[6,0] = 1 # v1->v7

         M_counts[1,6] = 1 # v7->v2
         M_counts[6,1] = 1 # v7->v7

         M_counts[6,3] = 1 # v4->v7
         M_counts[4,3] = 1 # v4->v5
         M_counts[3,4] = 1 # v5->v4

         M_counts[4,6] = 1 # v7->v5

         M_counts[5,5] = 1 # v6->v6

         M_counts[6,5] = 1
```

```
[[ 0.  0.  1.  0.  0.  1.  0.]
 [ 0.  0.  1.  0.  0.  0.  1.]
 [ 0.  0.  1.  0.  0.  0.  0.]
 [ 0.  0.  1.  0.  1.  0.  0.]
 [ 0.  0.  1.  1.  0.  0.  1.]
 [ 1.  0.  1.  0.  0.  1.  0.]
 [ 1.  1.  1.  1.  0.  1.  0.]]
```

Figure 6: Input relations between vertexes into matrix format

```
In [37]: def pagerank(M, d=0.85, square_error=1e-6):
             """
             M : the adjacency matrix of the pages. It is assumed to be column-stochas
         tic (ie column sum to 1); all links have equal weight.
             A page with no outgoing links (sink) is represented as a page with outgoi
         ng links to each other page (ie restart page).
             d: damping factor
             square_error : the algorithm iterates until the difference between two su
         ccessive PageRank vectors v is less than this (in squared norm)
             returns the PageRanks of all pages
             """
             n_vertexes = M.shape[0] # n_pages is the number of rows of M
             v = np.random.rand(n_vertexes) # initialize to random vector
             v = v / v.sum() # make v sum to 1
             last_v = np.ones((n_vertexes)) # will contain the previous v
             M_hat = d * M + (1-d)/n_vertexes * np.ones((n_vertexes, n_vertexes)) # eq
         uation (***) in Wikipedia page
             while np.square(v - last_v).sum() > square_error:
                 last_v = v
                 v = M_hat.dot(v) # at each iteration, progress one timestep
             return v


In [38]: pagerank(M)

Out[38]: array([ 0.041,  0.141,  0.024,  0.225,  0.217,  0.058,  0.293 ])
```

Figure 7: VertexRank calculation using d =0.85

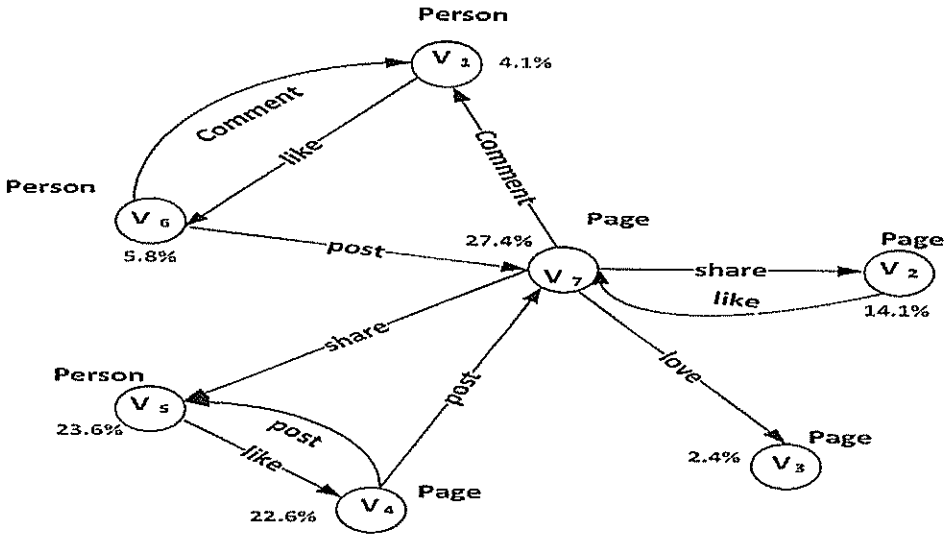

Figure 8: VertexRank result shows that v_7 has the highest rank

4. Conclusion

This paper presented keyplayers ranking analysis using social interactions network graph and centrality measurements. Betweenness and PageRank centrality are the two measurements utilized in this paper. A good keyplayer must have most direct route to another keyplayer and the betweenness centrality enable us to determine such keyplayer. A PageRank measurement can be used instead to determine how well a keyplayer connected to other well connected keyplayers. We have demonstrated using an example of social interactions network graph to determine betweenness measurement. The result shown that vertex v_7 has the highest betweenness measurement. This implies that vertex v_7 has the most direct route to other vertexes. We also implemented VertexRanking based on PageRank algorithm using python. The result shown that vertex v_7 is also has the ranking 27.4% follow by vertex v_6 and v_5 respectively. These three vertexes are the main keyplayers in the given SING. Given such information, we can venture that the three vertexes are good candidate for mining social interactions network data. Furthermore, if we are interested in promote a particular trend, these keyplayers might be the best candidate to help facilitate the promotion campaign. The work in this paper needs to be extended to a larger scale, moving toward big data. Now that a set of candidate keyplayers can be identified, we will extend our work toward investigating graph mining and text mining algorithms to extract rich semantic and useful graph structures.

5. References

D.R. White, S. P. B. (1994). Betweeness centrality measures for directed graphs. *Social Networks, 16*, 335-346.

Foster, P., Borgatti, S. P., & Jones, C. (2011). A gatekeeper is an indigenous member of the community who has either formal or informal influence with the culture. . *Poetics, 39*, 247-265.

Freeman, L. C. (1980). The gatekeeper, pair-dependency and structural centrality. *Quantity and Quality, 14*, 585-592.

Google. (2016). *Google Trend: FoodTrends 2016 US report.* Retrieved from https://think.storage.googleapis.com/docs/FoodTrends-2016.pdf

Grandjean, M. (2016). A social network analysis of Twitter: Mapping the digital humanities community. *Cogent Arts & Humanities, 3*(1).

Kieselmann, M. (2014). *What makes the difference? Investigating trendsetters' motivations.* (Master), Université Panthéon-Sorbonne - UFR d'Économie

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank Citation Ranking: Bringing Order to the Web.* Retrieved from http://ilpubs.stanford.edu:8090/422/

Tutte, W. T. (2001). *Graph Theory*: Cambridge University Press.

Zachary, W. W. (1977). An Information Flow Model for Conflict and Fission in Small Groups. *Journal of Anthropological Research, 33*(4), 452-473. Retrieved from http://www.jstor.org/stable/3629752

Zafarani, R., Abbasi, M. A., & Liu, H. (2014). *Social Media Mining*: Cambridge University Press.