Entrepreneurs’ Online Social Networks:
Structure and Characteristics

Okke Formsma

May 24, 2012

Master’s Thesis for Entrepreneurship M-track in Artificial Intelligence

Supervised by Tsvi Vinig, Yang Song
Abstract

The structure of social networks in which entrepreneurs engaged in are deemed critical to their performance. In this work we examine the network structure and characteristics of entrepreneurs on LinkedIn, Facebook and Twitter. Data was collected through a self-report survey that also collects the local neighborhoods for each entrepreneur on the aforementioned social networks.

We find that for entrepreneurs the degree distribution of LinkedIn and Facebook is not scale-free but rather exponential in nature. Earlier findings suggesting that entrepreneurs with a large network perform better than those with a small network could not be replicated. We do find that strong links (connections with high betweenness) on Twitter have a high chance to be present on the other two networks as well.
Acknowledgements

I would like to thank Yang Song and Dr. Tsvi Vinig for the guidance they provided and the valuable discussions we had over the past months. I would like to thank Yang Song specifically for access to her dataset, and Dr. Tsvi Vinig for introducing me to the field of networks in entrepreneurship. Finally, I would like to thank Paul Voskuilen for the brainstorms we had over coffee.
Contents

Abstract iii

1 Introduction 3

2 Prior Work 5
  2.1 Networks .......................................................... 5
    2.1.1 Nodes and Edges ........................................ 5
    2.1.2 Degree and neighborhood ................................ 6
    2.1.3 A note on drawing graphs ............................... 7
    2.1.4 Graph traversal ........................................... 7
    2.1.5 Random Graphs ............................................. 9
    2.1.6 Scale-free networks ...................................... 10
    2.1.7 Exponential networks ................................... 12
    2.1.8 Small worlds ............................................. 13
    2.1.9 Strong and weak ties .................................... 14
    2.1.10 Network of Networks .................................... 15
    2.1.11 Homophily .................................................. 15
    2.1.12 Graph partitioning and community structure ....... 16
  2.2 Social network analysis ........................................ 17
    2.2.1 Facebook .................................................... 18
    2.2.2 Twitter ...................................................... 18
  2.3 Entrepreneurial performance and networks ................. 19

3 Dataset 21
  3.1 Dataset .......................................................... 21
    3.1.1 Network overview ....................................... 21
    3.1.2 Profile matching ......................................... 26
    3.1.3 Future work ................................................ 30

4 Social Networks and Entrepreneurial Performance 33
4.1 Number of strong and weak links as indicator of performance ............... 33
   4.1.1 Hypothesis ................................................................. 33
   4.1.2 Method ................................................................. 34
   4.1.3 Results ................................................................. 34
   4.1.4 Conclusion ............................................................ 35
   4.1.5 Future work ........................................................... 35

4.2 Friends’ performance as indicator of performance ............................... 35
   4.2.1 Hypothesis ................................................................. 36
   4.2.2 Method ................................................................. 36
   4.2.3 Results ................................................................. 36
   4.2.4 Conclusion ............................................................ 37
   4.2.5 Future work ........................................................... 38

5 Structural analysis ............................................................. 39
   5.1 Degree distribution ...................................................... 39
      5.1.1 LinkedIn degree distribution ..................................... 39
      5.1.2 Facebook degree distribution .................................... 40
      5.1.3 Twitter degree distribution ...................................... 41
      5.1.4 Degree distribution over all three networks ................... 41
      5.1.5 Conclusion ............................................................ 42
   5.2 Network overlap ........................................................ 44
      5.2.1 Strong links hypothesis ........................................... 45
      5.3 Future work ............................................................. 46

6 Conclusions ................................................................. 49
Introduction

Online social networks (OSNs) are a widespread phenomenon; websites such as LinkedIn, Facebook and Twitter are ubiquitous. By explicitly listing the relationships between people, these websites give a unique insight in the social networks that exist in everyday life. This is not only interesting from a sociological viewpoint, but has applications in the research on entrepreneurial performance as well.

Before the emergence of online social networks, the networks of entrepreneurs were painstakingly mapped by conducting interviews. OSNs have made this process much easier, making much larger network studies feasible.

Network structure research has recently seen an insurgence with the advent of scale-free networks. Scale-free networks have specific properties which are often found in online social networks, such as having a small diameter (the small world phenomenon) and containing nodes that have a very large number of connections. The creation of a scale-free network is assumed to be preferential: nodes with many connections have a higher probability to attract new connections than nodes with few connections, leading to a rich-get-richer phenomenon. If a network is not scale-free, this assumption may not hold, and other mechanisms may form the basis of being well-connected in a network.

Because there multiple online social network available, it is interesting to see if they all have the same global structure. Also, it is a question which network is more valuable for entrepreneurs.

In this research we want to compare the structure of three online social networks of entrepreneurs: LinkedIn, Facebook and Twitter. We look at whether the networks are indeed scale-free. We present a method to find the profiles across different networks that belong to the same person. The overlap between the networks is scrutinized; we find that strong links have a higher chance to exist on multiple networks.
Prior Work

In this chapter, the basics of graph theory and social network analysis are explained. Then, a number of prior results on the structure of online social networks is explored. The final section reviews work on the link between social networks of entrepreneurs and their performance.

2.1 Networks

Networks in general, and social networks specifically, can be naturally represented as graphs. As graphs form the basis of this research, let us examine them in some detail.

2.1.1 Nodes and Edges

A graph shows the relationships between items. Each item in a graph is called a node (in mathematical contexts, a node is often called a vertex). Two nodes may be connected through a link, which we call an edge. In figure 2.1 a simple network with 4 nodes is depicted, connected with 4 edges. Each of the nodes is drawn as a small circle, and the edges are drawn as lines connecting them.

![Figure 2.1: A graph consisting of 4 nodes connected by 4 edges.](image)

Node A is connected with node B and node D through two edges. Another way of saying that two nodes are connected through an edge is to say that these two nodes are neighbors. So, nodes A and D are neighbours, and so are B and C, C and D, and A and C. The graph can be represented as

\[ V = \langle A, B, C, D \rangle \]
\[ E = \langle \langle A, B \rangle, \langle B, C \rangle, \langle C, A \rangle, \langle C, D \rangle \rangle \]

The graph in figure 2.1 could represent the friendships between four people: Alice, Bob, Carol and Dave. Because the node B in the graph is neighbors with A and C, Bob is friends with Alice
and Carol. Dave has only one friend: Carol.

In the first graph, the edges between nodes are symmetrical. When node A is connected with node B, the reverse is also true: node B is connected with node A. Thus, in this instance, friendship is always reciprocal. Graphs where edges are always symmetrical are called undirected graphs.

In contrast, directed graphs allow edges which signify an asymmetric relationship. In such a graph, nodes point to other nodes: A has a relationship with B, but B does not have the same relationship with A. For example, Alice could be in love with Bob while Bob is not in love with Alice. A graph composed of directed edges is a directed graph. When drawing a graph, directed edges are usually drawn as arrows. A small graph of only two nodes with one edge is drawn in figure 2.2. It can also be represented mathematically as

\[
V = \langle A, B, C, D \rangle \\
E = \langle \langle A, B \rangle, \langle B, C \rangle, \langle C, D \rangle, \langle C, A \rangle \rangle
\]

Figure 2.2: A directed graph.

In this thesis, we will encounter two undirected graphs, depicting the social networks of LinkedIn and Facebook. Connections in these networks are reciprocal. In addition, we see one directed graph: the Twitter social network. On Twitter, people can follow anyone, without approval of the other. In chapter \( \chi \alpha \chi \hat{\chi} \) we will compare these networks more thoroughly.

2.1.2 Degree and neighborhood

The degree of a node is the number of edges associated with it. In the undirected network of figure 2.1, node D has degree 1, as it is connected with one other node: C. Node C has degree 3, as it is connected with all three other nodes. A and B both have degree 2. In a directed graph a distinction is made between indegree and outdegree. The indegree is the number of edges pointing towards the node, and the outdegree the number of edges pointing outward. In figure 2, A and B both have indegree 1 and outdegree 1, C has outdegree 2 and indegree 1, and D has indegree 1 and outdegree 0.
The neighborhood of a node is the set of nodes has a link to. The neighborhood of C in figure 2.1 consists of \( \langle A, B, C \rangle \). The neighborhood of C in figure 2.2 consists of only \( \langle A, D \rangle \).

2.1.3 A note on drawing graphs

When drawing a graph, it does not matter in what locations nodes are drawn. The four graphs in figure 2 all depict the same graph. The positions of the nodes, the length of the edges and edge crossings do not matter. Nodes are typically aligned such that the structure of the relations between the nodes can be easily gauged. For large networks it usually is not feasible to layout and draw all nodes by hand. Many algorithms are specifically designed to automatically draw a graph in a such a way that it can be visually inspected. Many are implemented in the Gephi package \[1\] and the networkx python package \[2\].

![Figure 2.3: Various alternate representations of the graph from figure 2.1](image)

2.1.4 Graph traversal

Graphs consist of two building blocks: edges and nodes. In a graph, nodes are considered to be uniform they are all exactly the same. They differ in only one dimension: the edges they have with other nodes. In a similar sense, every edge is similar it links exactly two nodes. The structures created with these identical building blocks can be wildly complex. The source of complexity can only rise from the configuration of connections. Thus, to understand the properties of graphs, we must examine how they are connected.

Paths

Consider figure 2.1. While nodes A and D are not neighbors, it is possible to trace a path from A to D through C. We can write this path as \( \langle A, C, D \rangle \). There is another possible path from A to D, which runs over B and C: \( \langle A, B, C, D \rangle \). The first path has a length of 2: the path runs over two
edges. The second path has length 3: it runs over three edges. In directed graphs, edges can only be traversed in order.

In social networks, path length is sometimes referred to as the number of handshakes to someone. We will examine this, and the related popular notion of six degrees of separation, which states that the path length between any two people on earth is no more than six, below.

**Connectivity**

When a path exists between all nodes in a graph, we say that the graph is connected. Directed networks are called strongly connected when a path exists in both directions for each tuple of nodes. A directed network is weakly connected if after replacing all edges with undirected edges, the resulting network is connected.

In many networks, connectivity is of utmost importance. For example, the internet can be seen as a graph with computers and routers as the nodes and physical network connections between as edges. It is vitally important that a path exists between your computer and your email server. If no path existed between the two devices, communication would not be possible. Similarly, the graph needs to be connected, as otherwise some parts of the internet would not be able to communicate.

The internet is by definition connected (if you are not connected to the internet you are not part of the internet, so the internet graph can not contain nodes which not connected). Yet, many graphs are not connected. For example, the graph of married persons will consist of many separate parts. A connected graph can be trivially made unconnected by adding a node without edges to its set of nodes.

A connected graph consists of one single connected component. When the graph is not connected, it can be divided in multiple connected components. A connected component consists of a subset of interconnected nodes. A graph can contain any number of connected components. A node without edges is by itself a connected component of size 1.

![Figure 2.4: A graph with a disconnected component.](image_url)
When a graph is not connected, but has one very large connected component, this component is called the giant component. We will later see that many social networks have one giant component which includes the vast majority of nodes.

2.1.5 Random Graphs

How do graphs form? This question was answered almost simultaneously by Edgar Gilbert [3] and Paul Erdos - Alfred Rényi [4] in 1959. The models the authors propose describe very similar classes of graphs, generated by randomly creating edges between nodes. Thus the name: random graphs.

In the Erdos-Rényi model, a probability space of graphs is denoted $G(n, M)$ where $n$ is the number of nodes in the graph and $M$ is the number of edges in the graph. The model describes a probability distribution over all possible graphs with $n$ nodes and $M$ connections.

![Figure 2.5: An instance of $G(30, 0.15)$.](image)

Gilbert's random graph model is similar to the Erdos-Rényi model. It is denoted $G(n, p)$ where $n$ is the number of nodes in the graph and every possible edge exists with the independent probability $p$.

The models are intuitively rather similar: For $MpN$ (in which $N$ is the number of all possible edges $n(n - 1)$) the models $G(n, M)$ and $G(n, p)$ are almost interchangeable [5]. The Erdos-Rényi model can be seen as a generative model, where an instance $G(n, M)$ is generated by taking some $G(n, M - 1)$ and adding a random edge. Gilbert's model describes a probability distribution over every possible graph, with diminishing probability for graphs that deviate from the average. In this respect, $G_{ER}(n, M)$ and $G_{G}(n, M/(n(n - 1)))$ are not interchangeable. While the expected
number of edges for $GG$ is $EG = n(n - 1) \cdot p$, an instance from $G_G$ may have any number of edges. The degree of $GG$ is distributed according to a Poisson distribution [6] (it is a realization of $N = n(n - 1)$ identically independent distributed random variables).

The degree distribution in $G_{ER}$ follows a Poisson distribution as well. Nodes with degree much deviating from the average are exceedingly rare. Thus, nodes with a very low degree, only connected to a few other nodes, would occur infrequently. Nodes with many times more connections than the average are a veritable anomaly.

2.1.6 Scale-free networks

Random graphs are very convenient for analysis. However, they only describe a subclass of all graphs. Albert-László Barabási and colleagues investigated the structure of the world wide web [7, 8]. They found that the degree distribution of the world wide web does not follow the expected Poisson distribution. Instead, they found a number of sites with a great number of incoming links. As we’ve seen in the previous chapter, this is a very unexpected result if the network were a random graph.

Barabási et al., were not the first to note networks in which some nodes have a much larger number of connections to other nodes than the average node. In 1965, Derek de Solla Price [9] examined the citation structure of scientific articles. He found that the number of times an article is cited does not follow a poisson distribution. Instead, while most papers are only cited once, some papers are cited numerous times.

In his 2002 book, Linked [10], Barabási describes the quest of to explain the causes of the topology of the world wide web. What causes some webpages to collect thousands of incoming links, while most of the pages only get a handful? Random networks clearly do not explain this phenomenon. In random networks, there is a characteristic size (degree) for nodes, equal to the average number of connections. In scale-free networks, this characteristic scale does not exist: there are nodes that have multiple orders of magnitude larger degree. How do scale-free networks form? And in what way do they differ from random networks?

Multiple algorithms have been proposed to explain how scale-free networks form. One of these algorithms is preferential attachment, was proposed by Barabási and Albert [7]. The algorithm starts with a small number of nodes. At each time step, a new node is added. Along with the node, m edges are created. The edges connect the newly added node to other nodes. The other
nodes are selected with a probability proportionate to their degree. A node with degree 2 is only a quarter as likely as being selected as a node with degree 8. This creates a rich-get-richer process: any node that gets ahead in degree, will attract new links at a rate higher than nodes with fewer links. Bollobas [11] has proven that these two steps, growing the network and choosing the edges through a rich-get-richer scheme, indeed create a network with scale-free properties.

Note that the way the scale-free network or graph is generated is very different from how to make a random graph. When creating a random graph, it is supposed that all nodes are known upfront. In a random graph, the existence of an edge between two nodes is completely independent of all other connections. The preferential attachment method requires the graph to grow over time.

Barabási and Albert were not the first to explain how power laws arise in the graphs degree distribution. They were again preceded by Dered de Solla Price, who proposed a very similar algorithm in 1976 [12] which he dubbed cumulative advantage. Its method is very similar.

Besides the preferential attachment process, other algorithms have been proposed to explain the rich-get-richer phenomenon. Kleinberg et al. [13] observed that webpages are not randomly connected with other parts of the network, but tend to cluster together (cyber communities, according to the authors). Websites tend to link to other sites on the same topic. Their model, starts out with a small graph, and grows it by adding a node at every time step. The added node is connected with some randomly selected nodes. More edges are added by copying (a fraction of) the edges of another node. This creates a node that rather similar to an already existing node. While no thorough mathematical analysis is performed on this model as on the preferential attachment model, experimental analysis show that the model also generates scale-free networks. These networks have slightly different properties however: neighbourhoods tend to be better connected.

In a scale-free network, the degree of the nodes follow a power law. An ordered sequence \( y = [y_1, y_2, \cdots, y_n] \) of real numbers, where \( y_1 \geq y_2 \geq \cdots \geq y_n \), is said to follow a power law if

\[
k = cy_k^{-\alpha}
\]  

(2.1)

where \( k \) is the rank of \( y_k \), \( c \) is a constant and \( \alpha \) the scaling index. When plotted on a log-log scale, the relationship between the rank \( k \) and \( y \) appears as a straight line of slope \(-\alpha\).
2.1.7 Exponential networks

Another class of networks are the exponential networks. Many real-world networks belong to this class, such as the Worldwide Marine Transportation Network [15], the North American Power Grid Network [16] and the email network at the Rovira i Virgili university in Spain [17].

While the degree distribution of scale-free networks follows a powerlaw, the degree distribution of exponential networks follows an exponential formula of the following form:

\[ k = c \cdot e^{-\lambda x} \]  

(2.2)

Networks exhibiting a exponential degree distribution can be explained by non-preferential growing of a network [18, p. 25]. The mechanism is uncomplicated: create an initial network. Every time step, attach new node with \( m \) random connections until the desired network size is reached. Deng et al. performed a numerical simulation to test this algorithm on data from the afore-mentioned real-world and find evidence for the correctness of this algorithm [19]. The main difference between generating a scale-free network and a exponential network lies in the selection of new connections to attach a new node to.

To generate a scale-free network, preferential attachment is required: new nodes prefer to attach to nodes with high degree. In contrast, an exponential network is generated through a non-preferential attachment process: each existing node has an equal probability to connect to the
newly created node.

2.1.8 Small worlds

In his famous experiment, Stanley Milgram assessed the social distance between two random people in the United States of America [20]. He was interested in the number of links in the chain of acquaintances that connect two people.

To assess the distance between two people, Milgram sent out 296 packages to various people in Nebraska and Boston. These 296 people formed the starting points in as many chains. These people were asked to forward the package to a stockholder in Sharon, Massachusetts, near Boston. Instead of looking up the address of the stockholder, they should forward the package to someone they know personally, and who may be closer in the social network than themselves. Participants were asked to send a postcard to Milgram, so he could track the progress of the packages. Sixty-four of the packages, 29%, reached the intended participant. The average length of such chains was 6.1. This lead to the popular notion of six degrees of separation, which implies that any two people on earth are separated by no more than six connections.

Duncan Watts and Steven Strogatz investigated the distance between two nodes on a ring lattice [21]. They rewired the edges in the graph with a random probability $p$. The higher $p$, the more closely it resembles a random network.

For different values of $p$, Watts and Strogatz measured the average path length between two nodes $L(p)$ and local clustering coefficient $C(p)$. In a small world as observed in networks of actors, the US power grid and the genes of the worm C. Elegans, the path length $L$ should be relatively low while the clustering coefficient $C$ should be high.

In a regular network, $L$ and $C$ are maximal. In a random network, the path length tends to $L(1) = O(\ln(N))$, where $N$ is the number of nodes in the network. This is much smaller than $L(0)$, the length of paths in a regular lattice. For the clustering coefficient $C$ a similar effect is seen; in a regular lattice $C$ is maximal; the closer it gets to a random network, the lower the clustering component gets.

Watts and Strogatz observed that the path length between nodes in the network drops very quickly with increasing $p$, while the clustering coefficient drops much slower.

The world is even smaller in scale-free networks. Cohen and Havlin show that instead of growing proportionally with $\ln(N)$, average path length between two nodes in a scale-free network is
proportional with $\ln(\ln(N))$ [6]. Growing $N$ by several orders of magnitude has little effect on the path length in a scale-free network.

We will see the small world properties of online social networks Facebook and Twitter in sections 2.2.1 and 2.2.2, and results for our own dataset on entrepreneurs in section 5.1.

2.1.9 Strong and weak ties

In his seminal 1972 paper [22], Granovetter makes a distinction between strong and weak ties between people. Tie is another word for connection. While in regular graph theory the existence of an edge is binary (two nodes are either connected or they are not), Granovetter identifies two types of ties: strong and weak ties. The strength of a tie is depends on the amount of time, the emotion intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie.

Groups of people who know each other well form a community in the graph. The persons within these communities typically have many and strong ties to other people in the same community. A tie with someone from another community is called a local bridge. A local bridge would necessarily be a weak link, as it would be hard for two people from different communities to spend the resources necessary to strengthen their relationship.

Bridges to other communities prove a very valuable information source. Granovetter researched the sources of information leading to people finding a new job. Frequency of contact is used as a indicator of tie strength. The results indicate that most of the respondents learned about their new job by someone who they spoke to fewer than twice a week (16.7% more often than twice a week, 55.6% less than twice a week but more than once a year, 27.8% fewer than once a year).

The strength of weak ties lies in the wealth of information available from them. Strong ties are often with people who are very similar (also see the section on homophily below), and those people use very similar information sources. Seeking a new job or learning of other opportunities requires access to new information, which can be gained from the weak ties in ones network.

David Krackhardt responds to Granovetter's views by underlining the strength of strong ties [23]. While weak ties provide access to a large and diverse pool of information and resources, strong ties are built on a basis of trust. People connected by strong ties are more inclined to help each other, as they have a more intimate relationship. Krackhardt shows that strong ties are
especially valuable in turbulent environments where change and uncertainty reign.

Wellman and Wortley found that strong ties between people in are the main source for support for families [24]. However, their work explicitly does not include any work-related ties.

Recent work focuses on link prediction. For example, Liben-Nowell and Kleinberg examine the formation of co-authorship links of scientific papers [25]. They find that many similarity metrics are reasonably similarity suited for link prediction.

2.1.10 Network of Networks

A network of people is rarely found in isolation. Groups of people are interconnected through many different layers of networks, each partially overlapping with others. For example, a researcher takes part in a network with his colleagues at the university, with his friends and with the people at his sports club. Each of these networks is different, and the edges between the nodes have a slightly different interpretation for each of the networks. The person bridges the distinct networks, and creates a path between persons in the research group and the sports club. This concept of overlapping networks is called the Network of Networks, as coined by Craven and Wellman to express the complexity of social networks encountered in a study of urban life [26].

Related to the concept of the Network of Networks is multiplicity. The relation between two people is seldom based on a single well-defined relationship. Instead, a wide gamut of interpersonal relationships is available to humans. A relationship between two people consists of any combination of these relations.

It is worth noting that most social network studies only examine one kind of relationship between people. They only examine one type of bond: coworkers, friends or business associates for example.

2.1.11 Homophily

Homophily is the principle that a contact between similar people occurs at a higher rate than among dissimilar people [27]. Similarity between people can be found in a number of dimensions, such as ethnicity, age, religion, education, occupation and gender.

Connections between two people in a network do not form randomly. People can be introduced through a common friend. In this case, we can examine the network and understand why the relation between them formed. When people who are not introduced meet, they are typically
at a place such as work, or another social gathering. At these places, they meet people who are similar to themselves.

Homophily is the result of two mechanisms: selection and social influence. Selection is the process of choosing your friends. This can be both explicitly, such as choosing your friends in high school, or more subtly: by choosing where to work and live also biases the potential pool of people to connect with. The second mechanism, social influence, describes the alignment of peoples characteristics to better match the characteristics with their networks. Selection shapes the social network, social influence shapes the people.

A more elaborate treatise on the subject of homophily can be found in [28].

2.1.12 Graph partitioning and community structure

In the discussion of weak and strong links, we have seen that groups of people are tightly connected through strong links. The members of such a group have many mutual connections. We call such a group a community. Communities are connected through weak links.

The Girvan and Newman method for community detection uses the notion of betweenness to divide a network into separate communities [29]. It is based on the observation that many shortest paths between nodes run through the weak links that connect communities. For example, if there are two communities connected through only one weak link, all shortest paths between members of the first group and members of the second group must include this one edge. The Girvan-Newman method iteratively removes the edge(s) with highest betweenness until no edges are left.

The betweenness of an edge is exactly the number of shortests paths between nodes in the network that cross it. It can be intuitively understood by thinking of it as (traffic) flow. If each node would be a town and edges roads, betweenness is the amount of traffic on each road if there was traffic between all towns.

Betweenness can be computed with the following algorithm:

- set the betweenness of each edge to 0
- for each pair of nodes in the network:
  - find the k shortest paths between the nodes
  - for each edge on the shortest path:
2.2. SOCIAL NETWORK ANALYSIS

- add $1/k$ to the betweenness of the edge

The basic Girvan-Newman method works according to the following algorithm:

- while there are edges in the graph:
  - compute the betweenness of all edges
  - remove the edges with highest betweenness

This method starts out with the graph, and then iteratively removes edges. The process splits up existing communities into smaller subcommunities until there are no edges in the graph. The end state of the graph is not very interesting, as it is always a graph without edges. Instead, the method can for example be stopped when a certain criterion is reached. If, for example, one wants to find the 5 largest communities, the method can be slightly modified to read as follows:

- while there are less than 5 connected components:
  - compute the betweenness of all edges
  - remove the edges with highest betweenness

2.2 Social network analysis

Social network analysis focuses on the patterns of relationships between people. Scott gives a succinct overview of the field before online social networks were found [30]. Analysis was initially done after questioning one person [31] or a few people about their interactions with others, resulting in a dataset of a few thousand ties at most. Computer based interactions have made it much easier to track who is connected with whom, as recent studies on mobile phone data [32] and online social networks [33] [34] show. These recent studies analyze networks that span up to millions of users and millions of connections.

Online social networks (sometimes named after their medium: social network sites or SNSs) are a relatively recent invention. SixDegrees.com, often described as the first social network, was only launched in 1997. Since then, many online social networks have been created. For a succinct history of early online social networks, see [35].

While no two SNSs are identical, they all share three common features. Firstly, a user claims a unique identity, which is presented to the world through a profile page. Depending on the network, the user can put text, images and video to shape their online identity. Secondly, they
can form connections with other people. These connections may be symmetrical or asymmetrical, and may require approval by the other party. Finally, the users can communicate by sending each other messages, privately or public.

The openness of networks varies wildly. On some networks, the default is to share with everyone. An example of this is Twitter, where profiles, friend lists and messages are by default accessible to the world. In contrast, LinkedIn only shows very limited profile information to people not in your network, and messages between people are private.

The properties of a number of online social networks have been a grateful research subject. Social networking sites enable social network research on an unprecedented scale. Offline social network research is severely limited in scale by practical considerations; a tradeoff between the number of people interviewed and the depth in which the data was gathered was inevitable [31]. Gathering data on an online social network is much simpler; instead of requiring people to manually list their entire network, you can just ask for access permission to their readily maintained list of friends.

The remainder of this section will describe some work on two social network sites: Facebook and Twitter.

2.2.1 Facebook

Facebook is the largest online social network with 726 million (7.26 x 10^8) active users and 68.7 billion (6.87 x 10^10) friendship links between them. The Facebook graph is not publicly available; the default privacy settings restrict access to the friends list of a user to his friends. Ugander et al. [34], with cooperation from Facebook, were able to analyze the complete Facebook user graph.

They unexpectedly found that the degree distribution in the Facebook network does not follow a power law; there is significant curvature in the log-log graph of the complementary cumulative distribution function. The network contains hubs, as a small fraction of the users has a degree much larger than the average user (the median degree is 99).

2.2.2 Twitter

Twitter is microblogging service launched in 2006. Twitter users share tweets, messages of up to 140 characters with their followers. Kwak et al. [33], performed a quantitative study on the structure of friendships and information diffusion within Twitter.
Data for this study was collected by performing many thousands of page requests per hour from the Twitter API. (Currently, this method of collecting data is severely crippled; instead of 20,000 requests per computer per hour, access to Twitter data is constrained to 350 requests per user per hour.) More than 40 million user profiles were collected, along with almost 1.5 billion (directed) relations.

Among many other things outside of the scope of this work, Kwak et al. found that the degree distribution fits a power-law distribution with exponent 2.276, which suggests that the Twitter network a scale-free network. In section 5.1.3 we repeat this experiment, and find a similar distribution.

2.3 Entrepreneurial performance and networks

Do strong and weak ties influence the performance of an entrepreneur differently? Jenssen and Koenig examined the influence of strong and weak ties on latent entrepreneurs [36]. Their hypothesis is that strong ties give access to motivation and finances, and weak ties give access to information. Tie strength is defined by the degree of friendship acquaintances are considered weak ties, friends and close friends are strong ties. Strong ties (good friendships) are characterized by a high level of trust which is required for motivational and financial resources to be shared. Weak ties (acquaintances) provide more non-redundant information than strong ties, because the local networks of two people with strong ties often overlap considerably; weak ties provide access to new information.

The data in Jenssen and Koenig's study consists of a survey of 100 people who tried to start a business. These people were divided in two groups: those who actually started a business and those who did not. Jenssen and Koenig then examined the correlation between the number of strong and weak ties in both groups. The result of their analysis is a slight indication that strong ties are often a better source of motivational and financial resources and weak ties a better source of information. However, they must conclude the differences are not large, and for finances there is almost no difference between strong and weak ties. Thus, the intensity of friendship between two people does not influence the value of the relationship to the entrepreneur. The entrepreneurs network gathered in this study was sparse. Although no indication is provided to the number of ties gathered, the free recall mechanism described in the paper makes more than a dozen ties
for each entrepreneur unlikely. In addition, only ties which actually provided access to resources were considered, not the entire network of each entrepreneur.

Nann et al. examine the network structure of German entrepreneurs on the XING online social network network [37]. They find that the more connections entrepreneurs have and the higher their betweenness is in the network, the more successful the entrepreneurs are. In particular, links with alumni of the university where the entrepreneurs studied prove valuable: the more embedded the entrepreneur is in this network, the better his or her performance is.

Brian Uzzi, provides more insight in the role of networks in entrepreneurship [38]. Uzzi argues that the structure of social ties among firms shape economic action by creating unique opportunities and access to those opportunities. Social ties are disregarded in market theory, as all information that is exchanged is captured in the price of goods and services. Uzzi states that embeddedness provides trust, fine-grained information transfer, and joint problem-solving arrangements. Analysis of New York apparel industry data confirms that companies with stronger ties have a higher survival rate [39]. However, when a company becomes too dependent on only a few others (it has a few very strong ties), the probability of failure again rises.

Raz and Gloor found that start-ups in Israel that are better connected with peers have a higher survival rate than those who have fewer connections [40]. This is in line with results from the biotechnology sector in Canada [41].
3.1 Dataset

In this research, we use a dataset collected by Yang [Yang paper citation] during 2011. The data consists of two parts: a questionnaire and social network data.

The questionnaire was filled in by 373 individuals, of which 286 indicated that they were entrepreneurs. Unless otherwise indicated, all analysis in this thesis only include the entrepreneurs.

First, the entrepreneurs were asked some personal information, such as gender, birthyear, and their email address. Then, they were asked to fill out details about their company. If they have more than one company, they choose one to focus on. The next questions all concern this company. The following data is collected: year of incorporation, number of employees when the company incorporated, current number of employees, revenue growth in the first the company was incorporated, average yearly revenue growth, last years revenue growth. The entrepreneurs had to choose from a number of possible ranges for the number of employees (1-5, 6-10, 11-20, 21-50, 51-100, 101-500, more than 500) and revenue growth (0%, 0%-10%, 11%-20%, 21%-30%, 31%-40%, 41%-50%, more than 50%).

The respondents were asked for what purpose they use LinkedIn, Facebook, Twitter and Hyves. The options were ‘I dont use this network’, ‘Personal’, ‘Business’, or ‘Both personal and business’.

In the second part of the survey, the online social network of each entrepreneur was collected. The entrepreneurs were asked to grant us access to their network information, so our script could save their LinkedIn, Facebook, Twitter and Hyves connections. The script only accessed their profiles once to gather the connections they have with other people. The exact data gathered is detailed in the following section on the four networks.

3.1.1 Network overview

The entrepreneurs were asked to provide us with login information on four online social networks. The social networks were chosen for their wide use in the Netherlands during the data collection period. With the exception of Hyves, they are all global networks, and de facto standards in their respective areas.
The use of the LinkedIn, Facebook and Hyves social networks is very similar. In each of these networks, edges between users are undirected. They are created by the following mechanism. One person requests a connection to another. Only if the other person accepts this request, an connection is formed between them. When one of the users wants to revoke the connection, he can do so without consent of the other.

On these three networks, users create connections on the networks to access information. A user can access much more information about the other person when they are connected. For example, the list of connections of a person is only accessible to people they are connected with on each of these three networks. Pictures, short messages, network activity and personal information are similarly hidden for the world by default. Only people that are close in the network may access them.

Twitter is different. As discussed in section 2.2.2, Twitters social graph is more open. The content users post are by default accessible to everyone. Edges are directed, and do not require mutual consent to form.

**LinkedIn**

LinkedIn is the major business-related social networking site in the world, with more than 150 million members\(^1\). The adoption rate is high in the Netherlands compared to other countries, as 30% of the Dutch population has an account\(^2\)\(^3\). The LinkedIn graph is undirected. Links between two people are called connections, and are formed by the mutual consent model described above.

Of the 286 entrepreneurs, 261 shared their LinkedIn connections with us. Thus, the adoption rate of LinkedIn among the entrepreneurs was an astounding 91.3%! The 261 profiles contained 95,076 connections, which gives an sizeable 364 average connections. The median degree is 249.

The profile data of each connection was stored. In total, we collected 119,872 unique LinkedIn profiles (this includes connections of non-entrepreneurs who filled in the survey). The profiles contain the users full name, a headline (usually a job title such as Sales Engineer), industry (such as telecommunications or law practice\(^4\)), location (such as Amsterdam Area, Netherlands), the number of connections, country_code (e.g. nl or de) and an url to the profile picture of the per-

---

1. [http://press.linkedin.com/about](http://press.linkedin.com/about)
3. [http://blog.linkedin.com/2010/01/05/linkedin-netherlands/](http://blog.linkedin.com/2010/01/05/linkedin-netherlands/)
Figure 3.1: The LinkedIn network between entrepreneurs. Light-shaded nodes do not have a LinkedIn account.

son. Only for entrepreneurs connection information is available; the other profiles do not list an indication of how many connections they have, or with whom.

Facebook

As we've seen in section 2.2.1, Facebook is the largest online social network ever with in excess of 700 million users worldwide. In the Netherlands Facebook is the second largest social networking site with a little over 6 million users\(^5\).

Facebook has grown from approximately 3.4 million users in December 2010\(^6\) to the 6 million at the time of writing (March 2012). If this growth rate will continue, Facebook will surpass Hyves in the near future and be the largest social network in the Netherlands.

Edges between nodes in the Facebook graph represent a “friendship” relationship between people. An edge is formed similarly to the way it is in LinkedIn: one person requests the friendship to be formed, and when the other person accepts this request the edge is created. When one person revokes the friendship, it is deleted from the graph.

188 entrepreneurs shared their Facebook connections with us. This is 65.7% of the total number of entrepreneurs (286), suggesting a much higher use of Facebook than the average Dutch person (of which slightly more than 30% have an account). This could be because of the relative low age of the entrepreneurs. The average number of friends is 272, while the median is 215.

\(^5\)http://www.socialbakers.com/facebook-statistics/
We stored each friends’ profile. It contains the person’s name, gender, country code (such as ‘nl_NL’ or ‘en_GB’) and an url to a profile picture. 59,365 unique facebook profiles were stored. Just as with LinkedIn, only the friends of the entrepreneurs in the survey are available. For most profiles, we have no indication of with whom they are connected.

Hyves

Hyves is a Dutch social network site. It currently is the largest Dutch social networking site when considering the number of accounts (at the time of writing 9.7 million\(^7\)). However, some blogs have claimed\(^8\) that Facebook has surpassed Hyves in the number of active users. The relationship model of Hyves is identical to Facebooks. Neighbours in the graph are refered to as vrienden, the Dutch word for friend.

Only 27 out of 286 entrepreneurs listed a hyves account. This is less then expected, as the adoption rate of Hyves is supposedly more than 60% of the Dutch population. It is possible that many of the entrepreneurs still have a Hyves account, but dont use it or do not consider it relevant to share with us. In the Metrics section below, we will see that only 3 of the entrepreneurs use their Hyves account for business purposes.

The average number of friends on the Hyves network in our sample is 161.7, while the median

\(^7\)http://www.hyves.nl/over/
\(^8\)http://www.marketingfacts.nl/berichten/20110705_facebook_iets_groter_dan_hyves_in_nederland
is 104. These numbers are significantly lower than for the Facebook and LinkedIn graphs. Because of the small sample size, we have decided to not analyse the structure of the Hyves network in later chapters. An overview of the Hyves data is included in the Metrics section below.

We stored 8,758 Hyves profiles. Each contains one person’s name, gender, birthday, current age, number of friends, language (such as nl_NL), city (such as Amsterdam or Utrecht), country (such as Netherlands) and an url to a profile picture.

Twitter

Twitter is different from the three social networks discussed so far. It is branded as a “microblogging service”, instead of a “social network site”. Users of Twitter read and post messages of up to 140 characters. By default, these tweets are readable for everyone. A user can follow any other user, without permission of the other. A user sees a list of recent tweets by accounts he follows. This changes the interaction between users on the social network. As Kwak notices [33], people follow others not only for social networking, but for information, as the act of following represents the desire to receive all tweets by the person.

There is no official number of Twitter users. Kwak downloaded 40 million Twitter accounts in 2009 (also see section 2.2.2) [33], while CNN reports\(^9\) 300 million users in 2011. The adoption of Twitter by the Dutch is highest in the world\(^10\): 26% of Dutch people who use internet visited Twitter at least once in March 2011.

Of the 286 entrepreneurs in our dataset, 174 shared their Twitter information with us. The indegree and outdegree of twitter accounts is not symmetrical. The entrepreneurs followed on average 396 others, while the median lies on 174.5. In comparison, they were followed by 628 people with a median of 238. The large difference between average and median friends hints to a number of outliers which have a much higher than average number of friends. In section Metrics, below, we will examine this further.

We collected 114,907 Twitter profiles. Each profile contains a full name, username (a unique nickname, prefixed with @ such as @OkkeF), description (a short personal message such as A lonesome do-gooder with excellent communication skills), location (free-form input, ranging from


\(^10\)http://www.comscore.com/Press_Events/Press_Releases/2011/4/The_Netherlands_Ranks_number_one_Worldwide_in_Penetration_for_Twitter_and_LinkedIn
Figure 3.3: The Twitter network between entrepreneurs

Universe to Amsterdam), timezone (one of a number of predefined locations, such as Amsterdam, or Arizona) and an url to a profile picture.

3.1.2 Profile matching

For each of the entrepreneurs, we know which Facebook, LinkedIn and Twitter accounts correspond to the same person. The vast majority of entrepreneurs (73%) has an account on two of the three networks, while almost half (45%) has an account on LinkedIn, Facebook and Twitter. Only one entrepreneur did not fill in any social network account information in the survey. However, for most of the profiles in our database, the match between individuals and their social network profiles is not explicit. There is no information in the social media profiles to uniquely identify an individual. Matching profiles across different social media to find which belong to the same natural person is a duplicate record detection task. Duplicate record detection is an active research area [42, 43].

Because of the limited number of entrepreneurs that shared their Hyves profile with us (27), we have not included the Hyves network in this analysis. Similarly, in the remainder of this work, we will not include Hyves data unless otherwise indicated; only the data from the Facebook, LinkedIn and Twitter networks are included.

The “name” field is the only one which is the same across the three networks. Thus, a straightforward way of matching profiles is to compare full names of each profile, and mark them as belonging to the same person if the names are identical. This approach has an obvious drawback:
many people have the same name. For example, we found many people with the family names “Jansen”, “de Jong”, and “Bakker” and first names such as “Mark”, “Jeroen”, and “Peter”.

It is not possible to disambiguate between two people by only their names. However, we can utilize the graph of relations between people to find which profiles belong to the same person. Assume that social networks overlap significantly for each person: there is significant overlap between the set of people in a Facebook and LinkedIn neighborhood of the same person. Thus, to check if two profiles which have the same name belong to the same person, we can check the overlap in the network of the two profiles.

For example, see figure 3.4. We see two LinkedIn profiles (represented by yellow circles) with the name “Peter” and one Facebook profile (represented by a blue circle). The white circles are people with connections to one of the profiles. Which of the Peters are probably the same person?

![Figure 3.4: Profile merging example.](image)

Let’s see how much their profiles overlap. The Peter on LinkedIn (1) and Peter on Facebook profiles have 1 connection in common, while Peter on LinkedIn (2) and Peter on Facebook share 2. Because the Peter on Facebook and Peter on LinkedIn (2) profiles have the most connections in common, they most probably belong to the same person and we match the profiles.
Entrepreneurs | Non-Entrepreneurs
---|---
Before | After | Before | After
No Network | 1 | 1 | 0 | 0
LinkedIn | 261 | 261 | 119,611 | 119,611
Facebook | 188 | 197 | 59,176 | 59,167
Twitter | 174 | 179 | 114,733 | 114,728
LinkedIn and Facebook | 176 | 185 | 40 | 14,605
LinkedIn and Twitter | 158 | 163 | 25 | 5,564
Facebook and Twitter | 133 | 142 | 26 | 3,316
LinkedIn, Facebook and Twitter | 129 | 138 | 17 | 2,442

Table 3.1: Number of people in our dataset who have an account on one or a combination of social networks, before and after the matching procedure outlined in section 3.1.2. Before matching, different social media profiles belonging to one person were counted as distinct people. After matching, these profiles are associated with one person.

Matching results

Before the matching procedure, the database contains 294,144 social network profiles, of which 623 belong to entrepreneurs. We performed a cross-match as described in the previous section.

In table 3.1 the effects of the matching procedure on our dataset is shown. Let’s examine the results for the entrepreneurs (leftmost two columns and figure 3.5) first.

Before matching (using the raw data from the survey), 261 entrepreneurs are associated with LinkedIn, 188 with Facebook and 174 with Twitter accounts. 176 entrepreneurs are associated with both a LinkedIn and Facebook account, and 129 had listed accounts on all three networks. One entrepreneur indicated he did not have any social network profiles. During the matching process, 9 additional Facebook profiles were associated with entrepreneurs, as well as 5 Twitter accounts. Below (in section 3.1.2) we will discuss this further.

Besides the profiles that are associated with the entrepreneurs, there are 119,611 LinkedIn profiles, 59,167 Facebook profiles and 114,728 Twitter profiles in the dataset. All these profiles have first-degree links with at least one entrepreneur. After the matching procedure, there are 14,565 LinkedIn and Facebook known to belong to the same person, as well as 5,564 LinkedIn and Twitter profiles, and 3,316 Facebook and Twitter profiles. For 2,442 people an account was found on
3.1. DATASET

Figure 3.5: The size of the circles in this Venn diagram shows the number of profiles on each network. The overlap shows the people who have accounts on multiple networks. The merging procedure changed the overlap only slightly; very few profiles were found that were not listed in the survey.

each of the three networks.

Unexpected profile matches

The matching procedure found 9 Facebook accounts for Entrepreneurs who did not list one in the survey. Manual verification of additional profile information, such as the profile pictures, showed that the Facebook accounts probably did belong to the Entrepreneurs. This is remarkable, because all of them listed I dont use facebook on the For what purposes do you use facebook question. Similarly, 5 Twitter accounts were found for people who listed I dont use Twitter in the survey. In contrast, we did not find any LinkedIn profile for people who listed I dont use LinkedIn. One explanation for this is that the survey was conducted over a course of 9 months. Entrepreneurs who did not have Facebook when they filled out the survey might have created an account later. The adoption of LinkedIn in our target group is over 90%, while the adoption rates for Facebook and Twitter are 65% and 60%, respectively. Additional LinkedIn profiles were not found because almost all entrepreneurs already had LinkedIn when the survey was started. The matching procedure found those entrepreneurs who created a profile on the Facebook or Twitter pages after they filled in the survey. Another possibility is that these entrepreneurs chose not to share their Facebook or Twitter accounts because of e.g. privacy considerations.

30 Twitter accounts were found for entrepreneurs who already listed another Twitter account
Figure 3.6: Before the merging procedure, only a few profiles were known to belong to the same person. After merging, we know which profiles belong to the same person on different networks. These diagrams are approximate scale.

as their own in the survey. Our matching algorithm found a second Twitter account belonging to these people. The reason for this overlap is that many entrepreneurs have a Twitter account for their company, as well as a separate personal account. The survey let the entrepreneurs list only one account, and many chose to provide their company account, not their personal account.

3.1.3 Future work

An even better profile matching method could use profile images. Manual exploration of the profile data showed that many entrepreneurs use the same profile image across different networks. While the images across networks are of different size and quality and (in the case of LinkedIn) watermarked, they are generated from the same base image. Relatively simple image comparison software should be able to find the identical images.

Another improvement could be made by allowing variation in spelling of names. Johnathan Doe may list his full name on LinkedIn, while being known as John Doe on Twitter. Many Chinese and other non-Western names are originally written family-name-first, but reversed for use in international context. These names could occur in different order on different social networks.

We assumed that entrepreneurs have no more than one account on each network. Our results show that this is a valid assumption for LinkedIn and Facebook. However, we have seen that entrepreneurs use multiple Twitter accounts; usually one on a personal title and one representing their company. Many personal Twitter accounts refer to the company accounts in their description.
attribute. Future research could explore patterns in the corporate and personal Twitter accounts of entrepreneurs.
The social networks in which entrepreneurs are embedded have influence on the performance of entrepreneurs. We have seen prior work on this topic in section 2.3. In this section we examine the predictive value of the structure of online social networks on entrepreneurial performance of Dutch entrepreneurs.

First, we will look at the influence of strong and weak links on entrepreneurial performance. Then, we will look at whether the performance of an entrepreneur is correlated with the performance of his friends.

4.1 Number of strong and weak links as indicator of performance

Which is a better indicator for entrepreneur performance: the number of strong links or weak links an entrepreneur has? To answer this question, we will first see whether strong links or weak links are correlated with entrepreneurial performance at all.

4.1.1 Hypothesis

In line with Granovetter’s ideas on the strength of weak ties (see section section 2.1.9), our first hypothesis focuses on the strength of weak ties. The more weak connections a person has, the more sources of information are potentially available to her. This might lead to additional opportunities to improve and grow her business.

H1. A positive correlation exists between the number of weak connections an entrepreneur has and his/her performance.

The second hypothesis focuses on strong ties. Strong ties, marked by a high level of trust, have great influence on the behavior of individuals according to Krackhardt’s 1992 rebuttal of Granovetter (see section 2.1.9). The more strong ties a person has, the more she can count on social support from the people in her social network. This social support might a positive effect on his/her performance as entrepreneur. This leads to competing hypothesis H2:

H2. A positive correlation exists between the number of strong connections an entrepreneur has and his/her performance.

In our dataset, the strength of connections between two people is not explicitly recorded. In
line with Granovetters four criteria for strong links (see section 2.1.9) we propose the following. A strong connection is a connection that exists on multiple online social networks. A weak connection is a connection only present in one of the networks. The rationale behind this is that it takes an effort to connect to someone on a social network. The stronger the connection between two people, the more probable it is that they have connected through multiple networks.

4.1.2 Method

For each entrepreneur, each connection in his neighborhood is labeled either strong or weak, according to the number of networks in which it is present: strong for connections in two or more networks, weak for connections only found in one. The assumption is that the time and effort required to build strong links will yield more. The number of weak and strong connection is counted for each entrepreneur.

Five metrics in our dataset are related to entrepreneurial performance. The number of employees when the company was started and the current number of employees are in discrete bins (1-5, 6-10, 11-20, 21-50, 51-100, 101-500, more than 500), coded from 1 through 7. The revenue growth percentage of the company is also measured in bins (0%, 1%-10%, 11%-20%, 21%-30%, 31%-40%, 41%-50%, more than 50%), coded from 1 through 7 also.

Only entrepreneurs who listed at least two networks in the survey are considered: 209 in total.

4.1.3 Results

The results of Spearmans rank correlation with p-values are shown in table 4.1. Most of the tests showed no evidence for correlation. Most p-values corresponding to the correlations have too high p-values (shown in parenthesis) to be statistically significant. There is a small (statistically significant with alpha < 0.5) correlation between the number of weak ties with with the current number of employees, but this correlation also exists between strong ties and the current number of employees.

The increase in weak and strong ties can be explained by the fact that the total degree (the sum of weak and strong ties) is also greater for entrepreneurs who have a larger number of employees. The Spearman rank correlation of total degree with current number of employees is 0.17 with a p-value of 0.1.
4.2. FRIENDS’ PERFORMANCE AS INDICATOR OF PERFORMANCE

<table>
<thead>
<tr>
<th></th>
<th>Weak ties</th>
<th>Strong Ties</th>
</tr>
</thead>
<tbody>
<tr>
<td>First year number of employees</td>
<td>0.04 (0.53)</td>
<td>0.05 (0.47)</td>
</tr>
<tr>
<td>Current number of employees</td>
<td>0.15 (0.03)</td>
<td>0.16 (0.02)</td>
</tr>
<tr>
<td>Average revenue growth</td>
<td>-0.00 (0.95)</td>
<td>0.07 (0.30)</td>
</tr>
<tr>
<td>First year revenue growth</td>
<td>0.03 (0.67)</td>
<td>0.11 (0.11)</td>
</tr>
<tr>
<td>Last year revenue growth</td>
<td>-0.06 (0.41)</td>
<td>0.02 (0.80)</td>
</tr>
</tbody>
</table>

Table 4.1: Spearman’s rank correlation between entrepreneurship performance indicators and the number of strong and weak links of entrepreneurs.

4.1.4 Conclusion

The results show that neither the number of weak nor the number of strong ties exhibit substantial correlation with the entrepreneurial performance indicators in our dataset. We must refute both hypothesis 1 and 2.

4.1.5 Future work

There are many other methods to define weak and strong connections. Our approach uses the multiplicity of a tie as proxy for connection strength. This might be misguided. A number of other methods to assess connection strength could be used such as neighborhood overlap or by measuring the number of interactions between on social networks.

4.2 Friends’ performance as indicator of performance

Do successful entrepreneurs connect with other successful entrepreneurs? Is success contagious? Brian Uzzi found that companies that are embedded in a network have higher survival chances than those who rely on the market to find suppliers and customers [38]. We extend this line of reasoning to the social networks of entrepreneurs. In accordance with the homophily arguments made in the prior work chapter, we expect successful entrepreneurs to connect to other successful entrepreneurs.
4.2.1 Hypothesis

Connections with other entrepreneurs yield new information and opportunities. However, we hypothesize that the quality of an entrepreneur's network is indicative of his own performance. Thus, an entrepreneur who connects with successful entrepreneurs is more likely to be successful than an entrepreneur who connects with less successful entrepreneurs.

H3: A positive correlation exists between the performance indicators of an entrepreneur and the average of performance indicators of entrepreneurs within his/her network.

4.2.2 Method

We will call all entrepreneurs who have at least one entrepreneur in their neighborhood the focal entrepreneurs. For each focal entrepreneur, the set of entrepreneurs he or she is connected with is gathered, and an average over these entrepreneurs is computed for five metrics: the number of employees when the company was started, current number of employees, annual revenue growth when the company was started, average annual revenue growth and revenue growth in the last year. We will call these averages the neighborhood average performance indicators for each entrepreneur.

The correlation between neighborhood average performance indicators and the entrepreneurs own performance indicators is computed using Spearman's rank correlation. The Spearman rank correlation does not assume a linear relationship between the two variables (as Pearson's r does assume), only a monotonic relationship.

4.2.3 Results

176 entrepreneurs were connected with at least one other entrepreneur in our dataset. The correlation between their performance indicators and the average of performance indicators in their local neighborhood is shown in table 4.2.

None of the revenue growth correlations are statistically significant. The ‘first year number of employees’ performance indicator is statistically significant with alpha = 0.5. The Spearman correlation coefficient for this value is 0.19; indicating a small correlation.

Figure 4.1 shows a scatterplot of the first year number of employees with the local neighborhood average of the same metric. A small random value has been added to the x-value of the
4.2. FRIENDS’ PERFORMANCE AS INDICATOR OF PERFORMANCE

<table>
<thead>
<tr>
<th>Performance Indicator</th>
<th>Correlation Coefficient</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>First year number of employees</td>
<td>0.19</td>
<td>0.01</td>
</tr>
<tr>
<td>Current number of employees</td>
<td>0.03</td>
<td>0.71</td>
</tr>
<tr>
<td>Average revenue growth</td>
<td>-0.00</td>
<td>0.97</td>
</tr>
<tr>
<td>First year revenue growth</td>
<td>-0.02</td>
<td>0.77</td>
</tr>
<tr>
<td>Last year revenue growth</td>
<td>-0.02</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Table 4.2: Spearmans rank correlation coefficient with the local neighborhood average (p-value in parenthesis).

points to make them overlap less. 160 of the 176 entrepreneurs had between 1 and 5 employees in their first year of operation. As a consequence, 127 entrepreneurs are only connected to entrepreneurs who had between 1 and 5 employees in their first year of operation.

![Figure 4.1: Scatter plot between the first-year number of employees of each entrepreneur and the average of their friends number of employees. The size of the bubbles indicates the number of datapoints at the same coordinates.](image)

While a small correlation exists, almost all data points lie on the same spot, which makes the correlation meaningless.

4.2.4 Conclusion

Most of the performance indicators for entrepreneurs are not correlated with the average performance indicators of entrepreneurs found in their neighborhood. The only performance indicator that showed a small correlation was heavily biased towards one value, making the correlation
meaningless. We therefore can not accept hypothesis H3.

It is possible that the lack of results in this chapter is due to a sampling bias. Most entrepreneurs are only connected with a small number of other entrepreneurs. Many entrepreneurs in our dataset probably have connections with others that are not present in our dataset.

4.2.5 Future work

Instead of only looking at the neighborhood graph of each entrepreneur, it might be interesting to look at larger clusters of entrepreneurs and whether the clusters have different characteristics, as done by Nann et al. [37].
Structural analysis

In this chapter we will take a look at the structure of the three social networks entrepreneurs take part in. First, we will examine if the networks are indeed scale-free. In addition, we will see if the combined network is also scale free. Finally, we will look at the overlap between the different networks.

5.1 Degree distribution

Many real world networks are scale-free. There are indications that social networks are scale-free as well. We expect to see a power law degree distribution in each of the social networks. As we’ve seen in chapter 2.1.6, the following relationship should (by approximation) hold:

\[ k = cy_k^{-\alpha}, \]  

(5.1)

where \( y = (y_1, y_2, \cdots, y_n) \) is an ordered sequence of node degrees, \( k \) is the rank of \( y_k \), and \( c \) and \( \alpha \) are constants. \( \alpha \) is the scaling index. The relationship between \( k \) and \( y_k \) appears as a straight line on a double-logarithmic plot.

On the linear scale, the best fitting linear function fit is plotted in red. On the doubly logarithmic scale, the best fitting scaling function is plotted, and on the semi-logarithmic scale, the best fitting exponential function is plotted. The rank-degree plot of the scale-free network indeed shows a scaling relationship: the best fitting function is a scaling function.

5.1.1 LinkedIn degree distribution

Figure 5.1 shows the degree-rank doubly-logarithmic plot of the LinkedIn network. The distribution of the number of connections in the LinkedIn network is, other than we expected, not scaling. The best (least-squared error) scaling function is plotted as a red line.

In contrast, examine the third plot in figure 5.1, which shows the semilog plot of the same data, and the best least-squares exponential function \( (c \exp^{-\lambda y_k}) \) describing the data plotted in red again. Except for a few outliers (one person with more than 3000 friends), the exponential function fits the data rather well. This is unexpected, as we would expect a scaling function to fit the data well!
5.1.2 Facebook degree distribution

Figure 5.2 shows CCDF plots for Facebook on linear, semi-logarithmic and log-log scale, respectively. The results are very similar as on the LinkedIn network. We are not the first to observe this behavior in the facebook graph; Ugander et al. also noticed that the degree distribution is not scaling [34].

Figure 5.1: The CCDF for entrepreneurs on the LinkedIn network on different axis scales. The dotted red line is the best fitting scaling function, the striped green line is the best fitting exponential function.

Figure 5.2: The CCDF for entrepreneurs on the Facebook network on different axis scales. The dotted red line is the best fitting scaling function, the striped green line is the best fitting exponential function. The straight line formed by the data in the semi-log plot indicates the distribution follows an exponential function.
5.1.3 Twitter degree distribution

Because edges in the Twitter network are directed, we must examine the indegree and outdegree of the network separately. Figure 5.3 and figure 5.4 show the in- and outdegree, respectively. The best fitting exponential function has a better fit than the best fitting scaling function in both cases.

![Figure 5.3: Twitter indegree distribution for entrepreneurs (number of followers). Both distributions seem to fit the distribution up to degree 1000. After that, the scaling function (red line) expects much larger than observed degrees, and the exponential function (green line) underestimates the number of connections.](image)

For the whole twitter dataset (including non-entrepreneurs), the best fitting exponential function severely underestimates the long tail of the distribution, with many people have millions of followers and follow hundreds of thousands of others.

However, the full-network data is severely biased. There were approximately 300 million Twitter accounts at the time this data was collected. A person with 12 million followers thus reaches approximately 4% of the Twitter population. Alternatively, there is a 4% chance that a random person on Twitter follows this person. As we have collected connections of 176 people, the probability of including the highly-connected person in the dataset is is $1 - 0.96^{176}$, or $1 - 7.6 \cdot 10^{-4} \approx 0.9992$. Thus, the sampling biases our data towards people with large numbers of connections.

5.1.4 Degree distribution over all three networks

Of 129 entrepreneurs, we know their connections on all three networks. Figure 5.7 shows the outdegree distribution for entrepreneurs who have an account on every network. Connections on multiple networks are only counted once (e.g. the connection that exists between Ann and Bert...
Figure 5.4: Twitter outdegree distribution for entrepreneurs (number of people followed). Similar to the previous figure, the scaling function overestimates the degree of the best-connected entrepreneurs, while the exponential function underestimates this.

on LinkedIn and on Twitter only adds 1 to the degree of each).

The degree distribution over the sum of the networks clearly follows a exponential distribution.

5.1.5 Conclusion

Just as noticed in previous work outlined in section 2.2.2, we have seen that the degree distribution of the nodes in the social network does not follow a scale-free distribution. Recent work on specific social networks such as Twitter and Facebook shows a degree distribution different from the scale-free power law. Our results indicate that an exponential function fits the rank-degree function better than the powerlaw for the entrepreneurs.

Exponential networks can be generated through a non-preferential process (see section 2.1.7). The fact that the social networks are exponential indicates that the underlying process might also be non-preferential, or at least non-preferential with respect to the degree distribution of nodes.

Prior work on exponential networks (see section 2.1.7) indicate that exponential networks are often limited by geographical or sociological boundaries. For our social networks, this is not unexpected: an entrepreneur will connect to people he knows, not to random people who might have a strong network.

This might also explain the difference between the LinkedIn and Facebook networks on the
5.1. DEGREE DISTRIBUTION

(one hand and the Twitter network on the other hand. LinkedIn and Facebook are specifically designed to connect to people already known to you. In contrast, Twitter is often used more as a ‘news service’, in which the connections are formed with the most interesting news sources, not necessarily with the sources you personally know.

This leaves open the question of why some entrepreneurs in our network have a higher degree than others. Taking the non-preferential generative process as basis, one would expect that older nodes (entrepreneurs who joined a network relatively early) have a higher degree than others.

This question lies however outside the scope of this work.)
5.2 Network overlap

The entrepreneurs use the social networks both for business and personal purposes. The following tables shows the number of entrepreneurs who indicated in the survey to use a particular network for personal and/or business use.

While the main use of LinkedIn and Twitter seems to be business purposes, the opposite is true for Facebook. Most people who use LinkedIn and Twitter for personal purposes also use it for business purposes. Twitter is used for both purposes more than the other two networks, who cater more to a specific target group. For business purposes, LinkedIn is the largest network. For personal use, Facebook is most popular, while Twitter wins for business and personal use combined.

In section 3.1.2, we have seen that many entrepreneurs use more than one social network. Almost half has an account on all three of the networks. The table above shows that many use

<table>
<thead>
<tr>
<th>Network</th>
<th>Business</th>
<th>Personal</th>
<th>Both</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>LinkedIn</td>
<td>231</td>
<td>80</td>
<td>74</td>
<td>49</td>
</tr>
<tr>
<td>Facebook</td>
<td>90</td>
<td>169</td>
<td>77</td>
<td>104</td>
</tr>
<tr>
<td>Twitter</td>
<td>164</td>
<td>104</td>
<td>97</td>
<td>115</td>
</tr>
</tbody>
</table>

Table 5.1: The purposes of the social networks for each entrepreneur.
the networks for the same purpose as well; LinkedIn and Twitter for business, and Facebook and Twitter for personal purposes. Thus, we expect to see a large overlap of connections between LinkedIn and Twitter, and between Facebook and Twitter. Because the purposes for LinkedIn and Facebook are so different, we expect to find fewer connections in common over these two networks.

Figures A1, A2 and A3 in the appendix show the three different networks within the Entrepreneur community.

The overlap between the networks is calculated by dividing the number of connections on both networks by the number of connections in either network (the Jaccard similarity coefficient). Take the set of edges $A$ to be the connections in one network, and $B$ to be the set of edges in the other.

$$\text{overlap}_{A,B} = \frac{|A \cap B|}{|A \cup B|}$$ (5.2)

The overlap between connections between entrepreneurs on LinkedIn and Twitter is 21%, on Facebook and Twitter 19%, and on Facebook and LinkedIn 29%. Over the whole of the dataset, not limited to connections between just the entrepreneurs, the overlap is 2.5% on LinkedIn and Twitter, 2% on Facebook and Twitter, and 8.4% on Facebook and LinkedIn.

5.2.1 Strong links hypothesis

We expect more strong links to be present on Facebook and LinkedIn than on Twitter. To test this hypothesis, we iteratively remove the weakest links from the Twitter network, and see how many connections overlap with each of the other networks.

The weakest connection is defined as the edge with the highest betweenness centrality. The betweenness centrality of an edge is a measure for how important a certain edge is to keep the network connected and is commonly used for graph partitioning purposes. The edge with the highest betweenness centrality is removed, and the betweenness centrality for all edges is again calculated. This process disconnects the large connected components in the graph quickly. See section 2.1.12 for a detailed description of the used graph partitioning algorithm.

The connections which are removed in this way are the weakest. Well-connected cliques stay connected longest, while bridges between different groups of people are eliminated.

These two graphs show that after removing edges from the Twitter graph that have the highest
betweenness centrality, edges that survive in the graph have a higher chance of also being in the LinkedIn or Facebook graphs.

Clusters in the Twitter graph are thus likely to be found in the Facebook and LinkedIn graphs as well.

5.3 Future work

There is much to explore in the overlap of social networks. It would be interesting to find a model which predicts on which networks a relationship between two people or entrepreneurs is formed, based on their characteristics and existing network connections.

It would also be interesting to understand what processes underly the observed degree distributions of the different social media. The degree distribution on LinkedIn and Facebook might be explained through sociological means. After all, the unit of analysis is a person.
Figure 5.9: Graph: 23% of the edges on Twitter overlapped with Facebook edges. Iterative removal of edges with the highest betweenness centrality raises the percentage of Twitter edges that overlap with Facebook.
Conclusions

A method was presented to identify which profiles belong to the same person across different social media, based on overlapping neighborhoods. This method was employed on a dataset consisting of LinkedIn, Facebook and Twitter profiles.

We have found that the degree distribution of LinkedIn and Facebook is not scale-free. Instead, we found a exponential degree distribution in our dataset. The degree distribution of the Twitter network seems to be neither scale-free nor exponential in nature. The exponential degree distribution of the LinkedIn and Facebook dataset indicate that the process underlying the network creation is non-preferential with respect to the degree distribution: many connections may not be the driving force in attracting new links.

Significant overlap between the LinkedIn, Facebook and Twitter networks was found. We were able to identify people across the three different networks, and found that their local neighborhoods overlap across networks. Strong connections, identified through low betweenness, are more probably to be found across multiple online social media than weak connections are.
Bibliography


