

Beskrivelse og verdivurdering av sosiale nettverkstjenester

Martin Falck-Ytter

Master i kommunikasjonsteknologi
Oppgaven levert: Juni 2011
Hovedveileder: Harald Øverby, ITEM

Abstract

With their presence, Social Networking Services (SNS) introduced new services instantly available for worldwide consumption. During the last decade, the popularity of SNS has risen tremendously. Today, SNS have millions of users and create a large proportion of total worldwide web traffic. This has not remained unnoticed by businesses, which have increased their focus against this new market.

Several network laws have been proposed to model either user behavior or network value. However, the validity of the behavior laws has to a small extent been verified for SNS. Similarly, most common network valuation laws are based on a theoretical approach. It is therefore unclear how precise they are for SNS valuation. In this study, empirical findings are presented to clarify user behavior in SNS further and give more precise SNS valuation estimates.

The data used for analysis in this study were obtained from SNS themselves or by web-pages including relevant statistics.

The results in this study showed that Zipf's law could not be accurately fitted with popularity of Twitter members. Popularity of Youtube videos could to a large extent be accurately fitted with Zipf's law. Average content productivity increases with network size for SNS studied. The power response surface, as a function of network size and average content created per day, $\bar{V}_{prs}(n, c) = 14.1514 \times n^{0.892437} \times c^{0.167022}$ was the best model for SNS valuation. Using previous results in this study, the power response surface was converted to a function only dependent on network size. The proposed model then grew $n^{1.226481}$ in asymptotic terms - approximately as Tilly-Odlyzko's law.

Acknowledgement

I wish to thank associate professor Harald Øverby for thoughtful suggestions and feedback during the work with this thesis.

I would also like to thank to professor Bjarne E. Helvik for regression analysis suggestions and Cecilie Falck-Ytter for some last corrections.

Contents

1	Introduction	14
1.1	Background	14
1.2	Motivation	14
1.3	Problems	14
1.4	Limitations	15
1.5	Contributions	15
1.6	Organization of report	16
2	Network effect	18
3	Network laws	20
3.1	Sarnoff's law	20
3.2	Metcalf's law	20
3.3	Reed's law	22
3.4	Tilly-Odlyzko's law	23
3.5	Beckstrom's law	24
3.6	Zipf's law	25
3.7	Participation inequality and the 1% rule	27
3.8	Comparison of network laws	29
4	Findings from social networking services	34
4.1	Zipf's law and Twitter	36
4.2	Discussion on Zipf's law and Twitter	43
4.3	Zipf's law and Youtube	44
4.4	Discussion on Zipf's law and Youtube	51
4.5	Number of connections in a social network	52
4.6	Discussion on number of connections in a social network	52
4.7	Relationship between content created and size of social networking services	53
4.8	Discussion on relationship between content created and size of networks .	58
4.9	What is important for members of online communities?	60
4.10	Discussion on what creates value in a social network	61
4.11	Models for valuation of social networking services	62
4.12	Discussion on models for valuation of social networking services	71
5	Summary	74
6	Conclusions and further work	76

6.1	User behavior in social networking services	76
6.2	Valuation of social networking services	77
6.3	Further work	77
A	Request sent to various social networking services	82
A.1	Social networking services contacted	82
A.2	Email request	83
B	Paper	84

List of Figures

1	TV broadcast network applicable for Sarnoff's law	20
2	Potential connections in a telephone network with four members	21
3	Zipf probability mass function (logarithmic axes)	26
4	Cumulative Zipf distribution	27
5	The 1% rule	28
6	Comparison of network laws that only depends on the number of users (logarithmic axes)	30
7	Increase in network value after interconnection or merging with Tilly- Odlyzko's law as valuation	31
8	Increase in network value after interconnection or merging with Metcalfe's law as valuation	32
9	Increase in network value after interconnection or merging with Reed's law as valuation (logarithmic z-axis)	32
10	Bob Metcalfe's profile on Twitter	37
11	Number of followers for the 10 011 most popular twitter users	39
12	Fitting of Zipf's law and data from Twitter (logarithmic axes)	41
13	Fitting of Zipf's law and data from Twitter (linear axes)	42
14	Studentized residuals after fitting data from Twitter with Zipf's law	42
15	Occurrences of studentized residuals after fitting data from Twitter with Zipf's law	43
16	Video playback on Youtube	45
17	Number of views for the most popular Youtube videos	47
18	Fitting of Zipf's law and data from Youtube (logarithmic axes)	49
19	Fitting of Zipf's law and data from Youtube (linear axes)	49
20	Studentized residuals after fitting data from Youtube with Zipf's law	50
21	Occurrences of studentized residuals after fitting data from Youtube with Zipf's law	50
22	Comparison of quadratic, linear and power regression with actual data	54
23	Linear regression (logarithmic x-axis)	54
24	Studentized residuals from linear regression (logarithmic x-axis)	55
25	Occurrences of studentized residuals with linear regression	55
26	Quadratic regression (logarithmic x-axis)	56
27	Studentized residuals from quadratic regression (logarithmic x-axis)	56
28	Occurrences of studentized residuals with quadratic regression	57
29	Scatter of actual value, average content created and network size	63

30	Correlation between content created and market value (logarithmic axes)	64
31	Correlation between number of members and market value	64
32	Correlation between number of members and content created	65
33	Response surface for estimated network value when \bar{V}_{lrs} is used as valuation	66
34	Response surface for estimated network value when \bar{V}_{qrs} is used as valuation	67
35	Response surface for estimated network value when $\bar{V}_{prs}(n, c)$ is used as valuation	69
36	Comparison of $\bar{V}_{prs}(n)$ with existing network laws (logarithmic axes) . . .	71

List of Tables

1	Comparison of some of the network laws introduced	29
2	Some examples of network value with Sarnoff, Tilly-Odlyzko and Metcalfe's law as valuation	30
3	Gain in network value	31
4	Regression fits for data from Twitter, sorted by R^2	39
5	Several regression fits for data from youtube, sorted by R^2	47
6	Social networking services with information about average number of connections and number of members	52
7	Websites with information about content created and number of members	53
8	Several regression fits for content created as a function of number of members	53
9	Comparison of sum-of-squares and degrees of freedom	57
10	Norwegian online communities in the study	60
11	Reasons why online community members stop using the social service or using it less	61
12	Size, content created and market value of social networking services . . .	63
13	Residuals and accuracy of the linear regression model	66
14	Size, content created and market value of social networking services . . .	68
15	Estimated network value in USD with \bar{V}_{grs} for some common network sizes	68
16	$\bar{V}_{prs}(n, c)$ applied to data	69
17	Estimated network value in USD with $\bar{V}_{prs}(n, c)$ for some common network sizes	70

Acronyms

API Application Programming Interface

GFN Group-Forming Network

HTML Hyper Text Mark Up Language

ISPs Internet Service Providers

MSE Mean Square Error

R² Coefficient of Determination

SNS Social Networking Services

SS_{err} Residual Sum of Squares

SS_{opt} Least Sum of Squares

SS_{reg} Explained Sum of Squares

SS_{tot} Total Sum of Squares

URL Uniform Resource Locator

USD United States Dollar

1 Introduction

1.1 Background

The increase in popularity of Social Networking Services (SNS) the last decade has not remained unnoticed. With their presence, SNS enabled new services such as sharing of media, event planning and creation of interest groups instantly available for worldwide consumption. Today, a significant proportion of total web traffic is generated by SNS. As a consequence if this, businesses have increased their focus against this new multi-billion dollar market.

1.2 Motivation

Several network laws have been proposed to model either user behavior or network value. Examples of these laws include Sarnoff's law for broadcast network valuation, Metcalfe's law for valuation of communication networks and Zipf's law for estimating popularity of content. However, the validity of the behavior laws has to a small extent been verified for SNS. Similarly, most common network valuation laws are based on a theoretical approach. It is therefore unclear how precise they are for SNS valuation.

1.3 Problems

- What generates value in a network is a disputed question. Metcalfe's law states that network value is equal to the number of potential connections. Reed's law is even more optimistic and express network value as the number of potential subgroups. Beckstrom's law, on the other hand, has another way to measure value, as the law uses utility surplus of all network members to calculate network value. Andrew Odlyzko and Benjamin Tilly suggest that the value of a user grows as $\log(n)$, which leads to a total network value of $n\log(n)$. There is clearly a disagreement on how to estimate network value.
- Is each network connection of equal value? Metcalfe and Reed's law assign an equal value to each network connection, while Beckstrom's and Tilly-Odlyzko's law assign different value for different network connections. Obviously, both approaches cannot be correct.

- The nature of networks differs in the way they function. For example, some networks require subscription fees, some offers seamless communication with other networks while other networks have advertisement. Is it likely that one network law can accurately describe the value for all networks?
- How do you test the accuracy of a network law? Does the law explain why some networks choose to interconnect or merge and why some networks do not?

1.4 Limitations

The fitting of Zipf's law and data from Twitter in chapter 4.1 is based on the 10 011 most popular Twitter users, and not from a uniform selection of all Twitter members. Similarly, in chapter 4.3, the fitting is based on the 160 most viewed Youtube videos. There is thus no basis to conclude on data outside the observation range.

The following limitations apply to chapter 4.7 and 4.11. Types of content created in SNS vary very much. In some SNS, the creation of content is a time consuming process. Examples of this could be creation of blogs and uploads of videos. In other networks, the creation of content is a simple process. Status updates on Twitter is an example of this (for more information about Twitter, see chapter 4.1). In chapter 4.7 and 4.11, different types of content are not differentiated. Another thing to notice is that some sites only provided day-to-day data for content created. These are used as estimations for average content created. A third limitation in these chapters, is that the data collected for both network size and average content created are not uniformly distributed in the observation interval.

The models for social network value in chapter 4.11 use only average content created per day and network size as independent variables. Since various SNS provide different types of services, what creates value varies correspondingly. However, since the model is required to be practical, some simplifications had to be made. Another limitation in this chapter was the few observations available, as only six SNS provided the information needed. A third issue arose when network size, average content created and estimated value were not retrieved at the same date. This is dealt with as described in chapter 4.11.

1.5 Contributions

In this study, empirical data regarding user behavior in SNS are presented. Adjusting the exponent in the Zipf probability mass function, the best-fit function for popularity

of Twitter members and Youtube videos are calculated. Whether content productivity increases with network size for SNS studied is concluded. A response surface model for SNS valuation is presented and compared with existing network laws.

1.6 Organization of report

The phenomenon network effect is introduced in chapter 2. This phenomenon has an important impact on how to model network value of networks exhibiting this effect.

Chapter 3 overviews the most common network laws proposed: Sarnoff's law, Metcalfe's law, Reed's law, Tilly-Odlyzko's law, Zipf's law and the 1% rule. These laws are compared in chapter 3.8.

Chapter 4 presents findings about user behavior in SNS and three models to valuate such networks. Useful information and methodology is presented before each result. User behavior findings in SNS are presented in the following five chapters:

- Chapter 4.1: Can popularity of Twitter members be modeled with Zipf's law?
- Chapter 4.3: Can popularity of Youtube videos be modeled with Zipf's law?
- Chapter 4.5: Examples of number of connections in SNS.
- Chapter 4.7: Does content productivity increase with network size?
- Chapter 4.9: What is important for members of online communities?

Each of these chapters is followed by a discussion of the current topic.

Models for valuation of SNS are presented in chapter 4.11 and discussed in chapter 4.12.

The problems introduced in chapter 1.3 are discussed in chapter 5.

Conclusions containing the most interesting findings in this study and further work are presented in chapter 6.

The following items are in the appendix:

- Appendix A: a list of SNS contacted and the request sent to these.
- Appendix B: a paper based on this study written by Harald Øverby and myself.

2 Network effect

Network effect (or network externality) is a phenomenon where the utility of consumption is affected by the number of other users using the same or compatible products [1]. This effect can both be positive or negative, depending on whether subscribers value the network more/less as the number of users increase/decrease. Network effects are often mistaken for economies of scale. We distinguish between supply side and demand side economies of scale. Supply side economies of scale refers to cost advantages obtained by a company due to expansion. Demand side economy of scale, on the other hand, is a synonym for network effect.

With a positive network effect, subscribers value the network more as the number of members increase. In such networks, being the only member is pointless, since the utility of a user relies on interactions with other members. A telephone network is an example of a network that exhibits a positive network effect. Since the value of a user in a telephone network is derived from being able to connect to other people, a large network is preferable over a smaller network. Similarly, the network becomes more valuable itself, as existing customers are able to connect to the new subscriber. The same effect occurs in SNS. Large SNS are attractive to prospects, as a lot of acquaintances probably also are members of the network. Equivalent, the acquaintances will also benefit if the prospect choose to join the network.

Negative network effects occur when more users make the network less valuable, typically because of congestion and competition of resources. In such networks, exclusiveness is preferable, since it means less congestion. Examples of such networks are frequent flying memberships and VIP-access clubs.

It is also possible for a network to exhibit both positive and negative network effects. The Internet, for example, is a network where the value of subscription increases with the number of possible services and interactions. In this network you prefer a lot of websites to be available. Your utility does also increase, as you are able to communicate with your friends and family through SNS and chat services. However, the value of being connected to the Internet decreases as more users are competing for the same physical resources. A lot of active users on the Internet will decrease your utility if it means lower download and upload speeds, overload on servers, package loss and so forth.

When a network effect is mentioned in the remaining parts of this thesis, it can be interpreted as a positive network effect.

3 Network laws

The following subchapters introduce the most common network laws proposed. This chapter concludes with a comparison of the different laws presented.

3.1 Sarnoff's law

Sarnoff's law is attributed to David Sarnoff, an American pioneer in radio and television. The law states that the value of a broadcast network, where the content is sent from one-to-many, is proportional to the number of subscribers. The reasoning behind this is that the bigger audience, the more you can charge for advertisements in the network. Examples of broadcast networks where the law is applicable include newspapers, radio and television networks. Figure 1 illustrates a TV broadcast network applicable for Sarnoff's law.

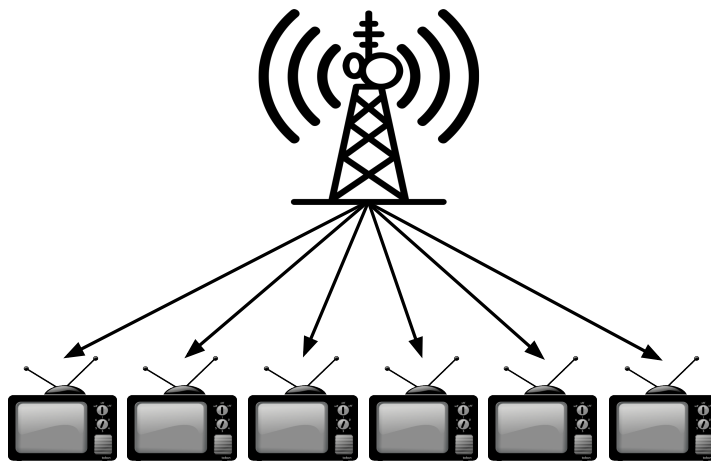


Figure 1: TV broadcast network applicable for Sarnoff's law

Sarnoff's law is widely accepted as valuation for broadcast networks, but also limited to this network type [2]. The law is given in equation 1:

$$S(n) = n \tag{1}$$

3.2 Metcalfe's law

Metcalfe's law states that the value of a network of n compatible communicating devices is equal to n^2 [3]. The law is applicable for one-to-one communication in a network of

n members. Examples of such networks include cellphone, instant messaging and email networks. The law can be understood mathematically as the number of possible links or unique connections in a network. In a network of n nodes, there are n nodes in the network that can reach the other $n-1$ nodes. This gives $n(n-1)$ links. But a link from a node A to node B in the network is the same as the link from node B to node A. Therefore; the total sum of unique links in the network is equal to:

$$M(n) = \frac{(n-1)n}{2}$$

$$M(n) \approx n^2 \tag{2}$$

As an example of potential connections in a network, consider the telephone network illustrated in figure 2.

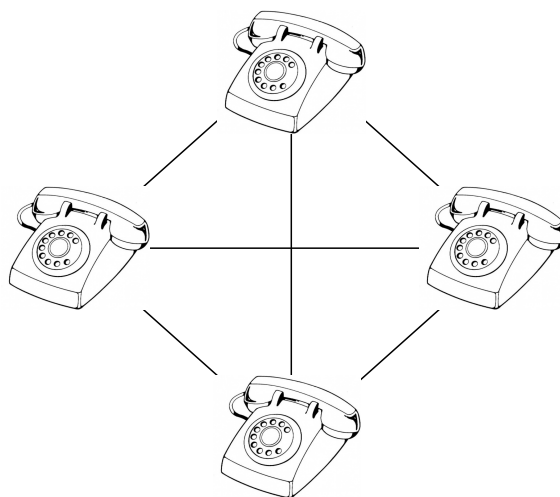


Figure 2: Potential connections in a telephone network with four members

The number of potential connections in this network is:

$$\frac{(4-1) \times 4}{2} = \frac{12}{2} = 6$$

As several papers have pointed out, among [2], [4] and [5], Metcalfe's law assumes that all network members are of equal value to each other. This can obviously not be true for all network sizes. Take the cellphone network in United States as an example. This network had an estimate of 285 610 580 subscribers in 2009 [6]. It is impossible that all users connected to this network will provide equal value to each other, if any value at all.

Aspects like culture, religion and geography affect the utility derived from connections in a network.

Andrew Odlyzko and Benjamin Tilly [2] also emphasized that Metcalfe’s law would provide incentive for all networks to merge or interconnect. According to Metcalfe’s law, two networks of size m and n will have a value of m^2 and n^2 respectively. If they interconnect or merge, the total value becomes $(m + n)^2$, which gives a surplus of mn for each network, or $2mn$ in total. Consider an example where two networks both have 1000 members. According to Metcalfe’s law, their value equals $1000^2 = 1\,000\,000$ separately or $2\,000\,000$ in total. If they interconnect or merge, the total value becomes $2000^2 = 4\,000\,000$, which means the network would be worth twice as much as the two separate networks. Such a "free lunch" would imply that all networks want to interconnect or merge. This is clearly not the case for many companies, as interconnections often require time and political pressure [2].

Robert Metcalfe replied to the criticism himself in a blog post and pointed out that the law was mostly applicable to smaller networks approaching critical mass [3]. He also argued that nobody had ever tried to estimate a , the constant of proportionality in his law ($M(n) = a \times n^2$). However, even if the constant of proportionality, a , is extremely small, Metcalfe’s law still grows $\Theta(n^2)$. Therefore, the term n^2 will dominate the function for sufficiently large values of n .

3.3 Reed’s law

In a paper from 1999, David R. Reed argues that there are some network structures where the value can scale even more than Sarnoff and Metcalfe’s law [7]. He introduces the concept Group-Forming Network (GFN) as a new network category that enables affiliations among subsets of members. Examples of such networks may be chat rooms and online auctions. Reed defines value as potential connectivity for transactions, which for a GFN is equal to the potential number of subgroups. In a network of n members, each element can be included or not in a subgroup. This gives 2^n possible subgroups in total. However, this includes two non-proper subsets: one where no elements are included and n sets where only one element is included. Therefore, according to Reed’s law, the value of a GFN is equal to:

$$\begin{aligned} R(n) &= 2^n - n - 1 \\ R(n) &\approx 2^n \end{aligned} \tag{3}$$

As equation 3 shows, Reed's law states that the value of such networks scale exponentially with network size. But what about networks where the value is derived from several types of communication categories? In such cases, Reed argues that the dominant component will out rule the least significant component(s) for sufficiently large values of n. So, if a network, for example, consists of components that scale accordingly to Sarnoff, Metcalfe and Reed's law, the component belonging to Reed's law will eventually dominate, since $2^n O(n)$ and $2^n O(n^2)$.

Since Reed's law grows even faster than Metcalfe's law, it is vulnerable for the same criticism. However, it is important to highlight that Reed talks about value of potential and not actual affiliations. This fact makes the law unpractical for real network valuation. To see this, consider how much value a new user increases the network value:

$$R(n + 1) - R(n) = 2^{n+1} - 2^n = 2^n(2 - 1) = 2^n$$

In other words, user $n + 1$ will always double the value of the network, which leads to an unrealistic growth in network value. To illustrate this, consider two networks with 100 members each. According to Reed's law they are separately worth $2^{100} = 1.2677 \times 10^{30}$ or 2.5353×10^{30} in total. If the networks interconnect or merge, the total value becomes $2^{200} = 1.6069 \times 10^{60}$. This would mean an increase in total network value of $6.3383 \times 10^{31}\%$.

3.4 Tilly-Odlyzko's law

In the paper "*A refutation of Metcalfe's Law and a better estimate for the value of networks and network*" [2], Andrew Odlyzko and Benjamin Tilly accuse Metcalfe and Reed's law for overestimating the value of networks. They argue that the main fundamental fallacy underlying Metcalfe and Reeds law is the assumptions that all potential connections or subgroups are of equal value to a network member. They reason that, since some connections are not used at all and some very rarely, an equal assignment of value to each connection or subgroup is not justifiable. They suggest a new way to value a general communication network of size n. Based on Zipf's law, Tilly and Odlyzko argue that a network participant, in a network of size n, derives value proportional to $\log(n)$. This leads to a total network value of:

$$T - O(n) = n \times \log(n) \tag{4}$$

This model has a growth rate only slightly faster than Sarnoff’s law. They argue that this is a better network law than Metcalfe and Reed’s law since:

- Their estimate provides only small gains in value when large firms interconnect, which explains why interconnection often requires time, effort and governmental regulatory.
- Large Internet Service Providers (ISPs) often refuse to exchange traffic freely with smaller ISPs without any payment. This is consistent with $\log(n)$ as valuation, since the smaller firm gains considerable more than the larger firm.

Even though Tilly-Odlyzko’s law seems to be able to describe real world observations of network effects, there are some downsides with the law. In their reasoning, Tilly and Odlyzko assumed that a network member derives value according to Zipf’s law. However, Zipf’s law is intended to describe popularity, not value. Whether this approximation is justifiable remains unclear. In addition, Odlyzko and Tilly did only provide some examples where Zipf’s law could be an accurate describer of popularity. Whether the law is a good estimation of popularity in all networks remain unanswered. As we later shall see, it is also important to estimate the exponential value in Zipf’s law. Without the exponential value specified, the function might differ very much; as the only restriction is that it is greater than 0.

3.5 Beckstrom’s law

In the paper *“A New Model for Network Valuation”* [8], Rod Beckstrom proposed a new model for network valuation. According to Beckstrom, the model can be used to value any network type and size. In this model, the present value of any network is equal to the sum of the net present value of the benefit of all transactions minus the net present value of the cost of all transactions. Note that transactions only are carried out if the benefit is higher than the cost of the transaction. All values are discounted over any given period of time. In mathematical notation, Beckstrom’s law is formulated as¹:

$$\sum_{i=1}^N V_{i,j} = \sum_{k=1}^M \frac{B_{i,k}}{(1+r_k)^{t_k}} - \sum_{l=1}^P \frac{C_{i,j}}{(1+r_l)^{t_l}} \quad (5)$$

Where:

¹The original paper [8] has some typos. In this study, r was changed to r_k in the formula and 1 to l under the explanation of t_k or t_l

$V_{i,j}$ = net present value of all transactions of $k = 1$ through n to individual i with respect to network j

i = one user of the network

j = identifies one network or network system

$B_{i,k}$ = the benefit value of transaction k to individual i

$C_{i,l}$ = the cost of transaction l to individual i

r_k and r_l = the discount rate of interest to the time of transaction k or l

t_k and t_l = the elapsed time in years to transaction k or l

Beckstrom defines benefit of a network transaction as difference between costs paid in the network minus the lowest cost alternative. A network transaction will not be executed if the network does not provide the lowest cost alternative. To illustrate the principle, consider the following example of cellphone subscription: you need to call your friend and have to pay 1\$ to your network operator for the entire conversation. As the next cheapest alternative, you can drive to your friend and talk to your friend in person. If we assume that this alternative cost 5\$, your benefit of subscribing to the network provider is equal to $5\$ - 1\$ = 4\$$ for this transaction. If we further assume that the cellphone provider has a cost of 0.10 \$ for your call, it means the provider has a benefit of the transaction of $1\$ - 0.10\$ = 0.90\$$. The total benefit value of this transaction, according to Beckstrom's law, is equal to $4\$ + 0.90\$ = 4.90\$$. If we discount each transaction with the appropriate discount rate raised to the time elapsed to the transaction, we get the net present value of the transaction. Finally, if we sum all the benefit value for all transactions in a network over a given period of time, we get the total network value for a given period.

Even though Beckstrom's law may give correct results, it introduces a new problem: How are you going to get the beneficial value and cost of *every* transaction in a network? This question must be as hard to answer as the original problem: How valuable is a network?

Others accuse Beckstrom for reinventing Metcalfe's law, as pointed out by [9] and [10]. In [10], the author derives Metcalfe's law from Beckstrom's law with simple algebra.

3.6 Zipf's law

Zipf's law is named after George Kingsley Zipf and refers to the fact that several types of data follow a Zipfian distribution. If k is the rank of elements from a data set (where $k = 1$ is the most frequent element), Zipf's law predicts that out of a population of N

elements, where s is the value of the exponent, the frequency of elements of rank k is:

$$f(k, s, N) = \frac{1/k^s}{\sum_{n=1}^N (1/n^s)}, s > 0 \in \mathbb{R}, n \in \mathbb{I} \quad (6)$$

Equation 6 is plotted in figure 3 with logarithmic axes. The figure shows the frequency of element $k = [1, 10]$ with $s = [1, 4]$ in a Zipfian distribution:

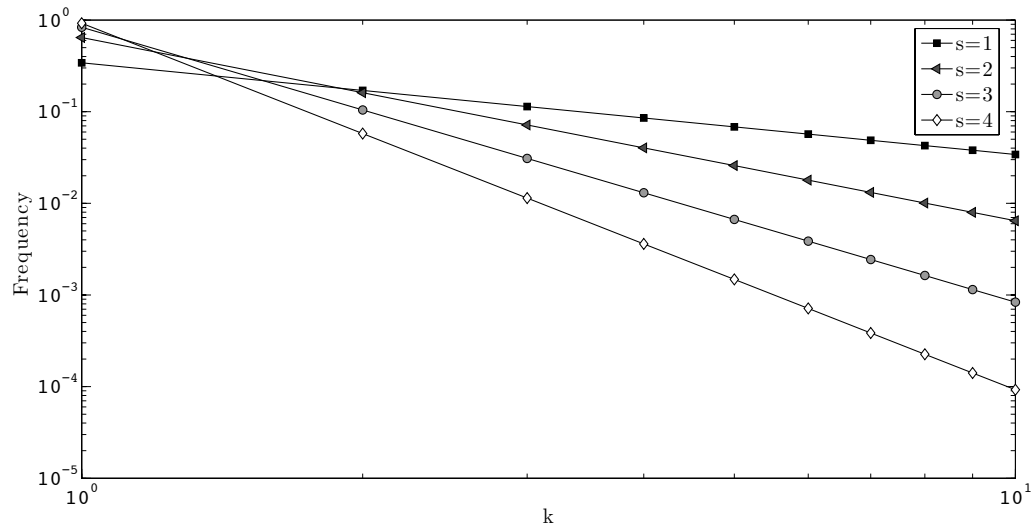


Figure 3: Zipf probability mass function (logarithmic axes)

The cumulative Zipfian function is plotted in figure 4 where $k = [1, 10]$ with $s = [1, 4]$. The cumulative frequency is always equal to 1 when all the elements in k are summarized.

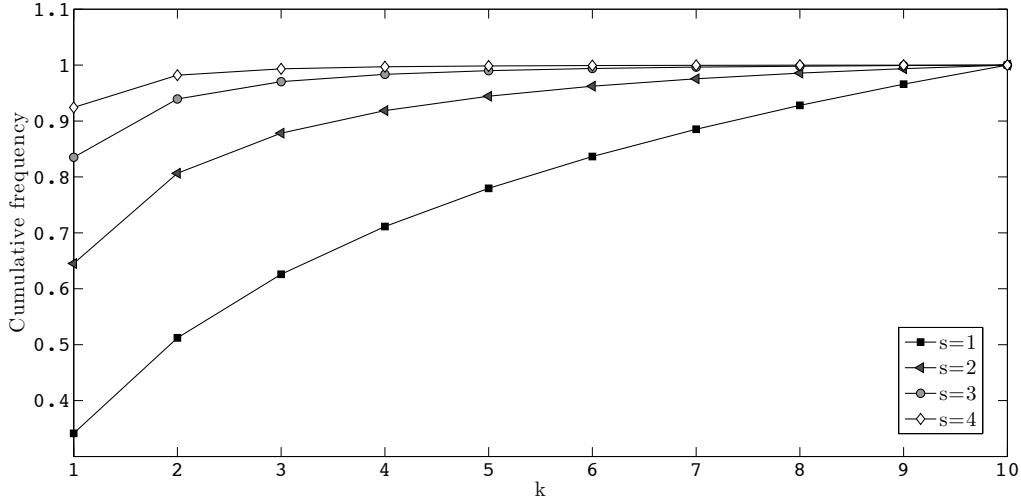


Figure 4: Cumulative Zipf distribution

Zipf's law has proven to be very accurate for modeling popularity of data, such as words in the English language [11] and sizes of large cities [12]. In *"Power Laws, Weblogs, and Inequality"*, Clay Shirky showed that income, web page links and traffic to sites follow a power law distribution [13]. In *"Zipf's law and the Internet"*, Lada A. Adamic and Bernardo A. Huberman shows that a great number of Internet features follow a Zipfian distribution [14]. In their research, they found Zipf's law to be present in:

- The level of routers transmitting data from one geographic location to another.
- The content of the World Wide Web.
- How individuals select the websites they visit and form peer-to-peer communities.

3.7 Participation inequality and the 1% rule

Participation inequality means that some people participate more than others. The phenomenon is well known and present in several situations in everyday life. A situation exemplifying the principle may be a conversation between coworkers. In this case, typically a few of the extrovert workers with the best subject knowledge talk a lot, while the majority talk little or nothing at all.

The 1% rule or the 90-9-1 principle divides a community into three categories: creators, editors and audience. The principle states that out of the content created in a community:

- 1% of the visitors will create content (creators).
- 9% will comment or modify (editors).
- The majority of 90% will just consume/read the content (audience).

The relationship between actors in a society following the 1% rule can be illustrated with a pyramid:

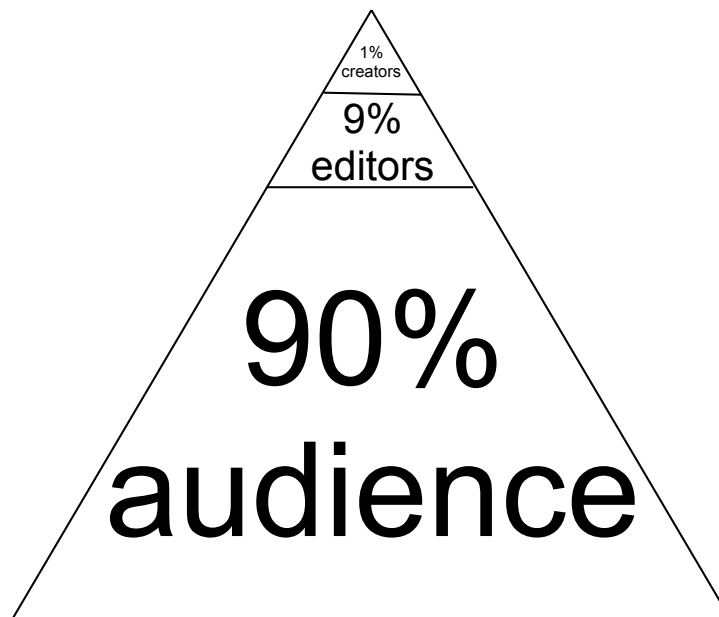


Figure 5: The 1% rule

The 1 % rule has been proved to be valid in several domains [15]:

- 167 113 of Amazon's book reviews were contributed by the top 100 reviewers.
- Over 50% of all the Wikipedia edits are done by 0.7% of the total users.
- In December 20, 2007 on the MSDN Community site, edits were made by 1.72% of the community.
- 0.16% of all visitors to YouTube upload videos to it.
- 0.2% of visitors to Flickr upload photos.

There are some downsides with the 1% rule. In situations where customer opinions is important, the 1% rule implies that a small share of customers give feedback. This gives an unrealistic picture of the customer base modeled. Similarly, if you try to find out what movies to watch or books to read, the 1% rule implies that most of the reviews written, represent a tiny share of people with experience about those items. Even though

this relationship seems to occur naturally, some means can be initiated to decrease the inequality. Participation rewarding and emphasizing the importance of contributions motivates users to actively participate. In addition, making the contribution process easier makes the threshold lower for contribution incentives.

Even though the 1% rule is present in several communities, the rule does not seem to be valid in all situations. In *"Crowdsourcing Participation Inequality: A SCOUT Model for the Enterprise Domain"*, Osamuyimen Stewart, David Lubensky and Juan M Huerta studied participation levels inside an enterprise network [16]. They claim that a 33-66-1 (33% audience, 66% editors and 1% creators) distribution can be achieved through careful design.

3.8 Comparison of network laws

Sarnoff, Metcalfe, Reed and Tilly-Odlyzko's law are all simple to use, but also limited to a specific network domain. Sarnoff's law is generally accepted for valuation of broadcast networks. Metcalfe and Reed's law on the other hand, talks about potential and not actual value, which leads to a heavy overestimate in network valuation as the number of members increase. Tilly-Odlyzko's law seems to be more accurate when it comes to describing real word examples of network interconnection or merging. A drawback with the law is that it is unclear what kind of value the law predicts. Beckstrom's law is applicable to all networks, but very little practical to use. Zipf's law, even though not a network valuation law, gives a handy description of how networks tend to function and a relationship between the most popular resources used. Another law not applicable for network valuation is the 1% rule. This law describes relationship between participation levels in communities where content is created.

Examples of network types applicable for some of the network laws presented are given in table 1. All network laws in the table find value for a single member or transaction in the network, and then sum for the total number of members in the network.

Law:	Applicable for	Examples of networks
Sarnoff	Broadcast networks	TV, radio
Metcalfe	Communication networks	Telephone, fax
Reed	Group affiliations networks	Online auctions, SNS
Tilly-Odlyzko	Communication networks	Telephone, fax
Beckstrom	All networks	TV, telephone, SNS

Table 1: Comparison of some of the network laws introduced

A comparison of Sarnoff, Metcalfe, Reed and Tilly-Odlyzko’s law is shown in figure 6.

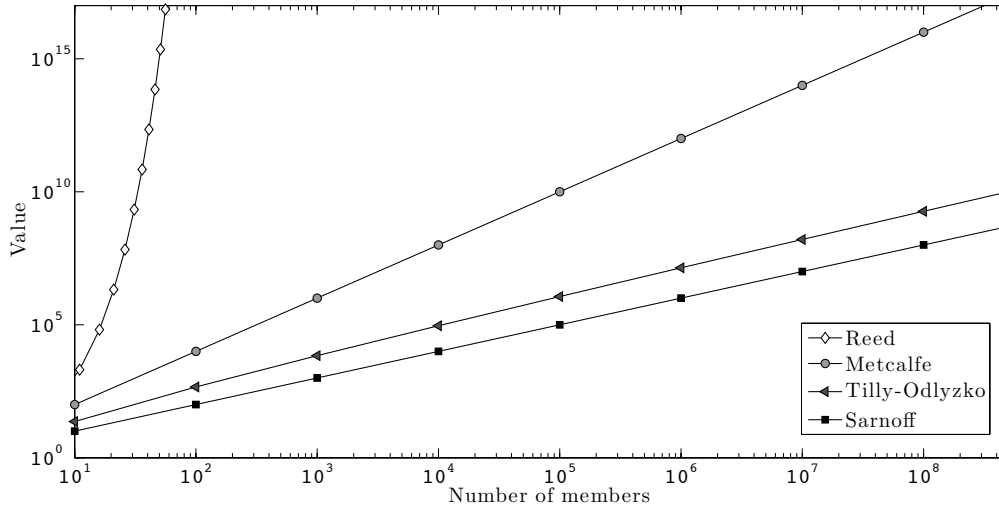


Figure 6: Comparison of network laws that only depends on the number of users (logarithmic axes)

Some numerical examples of network value according to Sarnoff’s, Tilly-Odlyzko’s and Metcalfe’s law are given in table 2. Reed’s law is left out, as the results are too large to be represented by most common math software.

Network size n	Sarnoff’s law $S(n) = n$	Tilly-Odlyzko’s law $T - O(n) = n \log(n)$	Metcalfe’s law $M(n) = n^2$
10 000	10 000	40 000	100 000 000
100 000	100 000	500 000	1 000 000 0000
1 000 000	1 000 000	6 000 000	1 000 000 000 000
10 000 000	10 000 000	70 000 000	100 000 000 000 000
100 000 000	100 000 000	800 000 000	10 000 000 000 000 000
1 000 000 000	1 000 000 000	9 000 000 000	1 000 000 000 000 000 000

Table 2: Some examples of network value with Sarnoff, Tilly-Odlyzko and Metcalfe’s law as valuation

The network laws proposed leads to different gain in value if networks interconnect or merge. Table 3 shows a comparison of gain in network value if network m and n interconnect or merge, according to the different laws.

Law:	Value of network m/n	Value of separate networks	Value of interconnection/merge
Sarnoff	m/n	$m + n$	$m + n$
Tilly-Odlyzko	$m \log(m) / n \log(n)$	$m \log(m) + n \log(n)$	$(m + n) \log(m + n)$
Metcalf	m^2 / n^2	$m^2 + n^2$	$m^2 + n^2 + 2mn$
Reed	$2^m / 2^n$	$2^m + 2^n$	2^{m+n}

Table 3: Gain in network value

The increase in network value if network m and n interconnects or merge, with Tilly-Odlyzko's law as valuation formula is plotted in figure 7. The increase is calculated as $(m + n) \log(m + n) - m \log(m) - n \log(n)$.

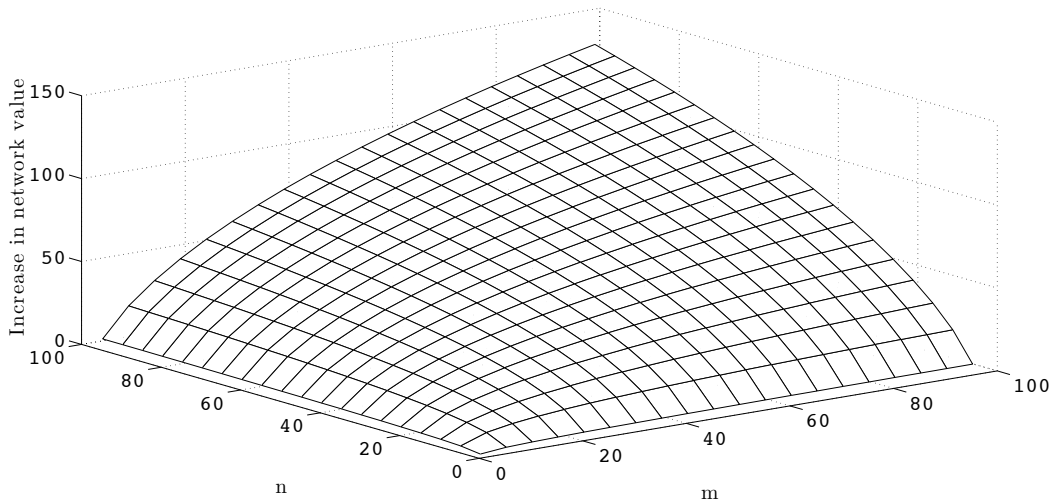


Figure 7: Increase in network value after interconnection or merging with Tilly-Odlyzko's law as valuation

The increase in network value if network m and n interconnects or merge, with Metcalfe's law as valuation is plotted in figure 8. The increase is calculated as $(m + n)^2 - m^2 - n^2$.

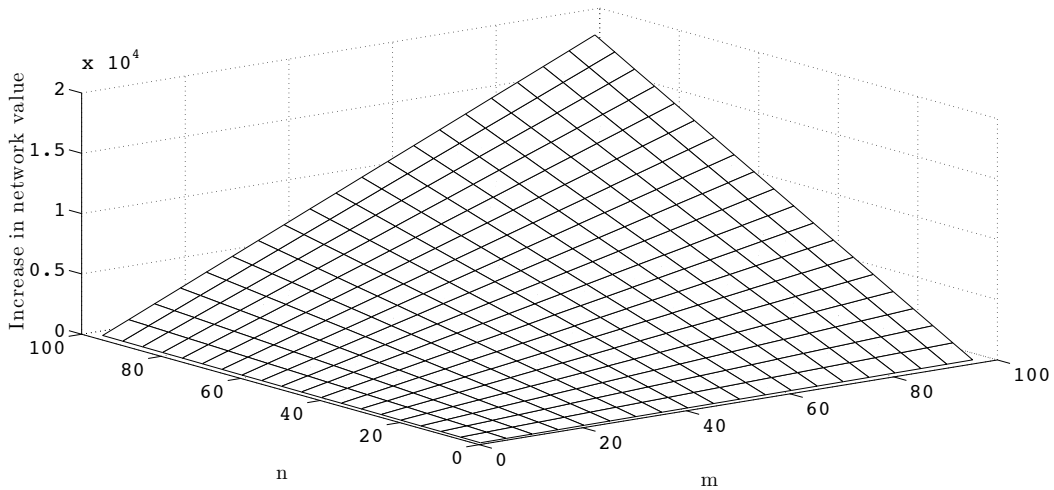


Figure 8: Increase in network value after interconnection or merging with Metcalfe's law as valuation

The increase in network value if network m and n interconnects or merge, with Reed's law as valuation is plotted in figure 9. The increase is calculated as $2^{m+n} - 2^m - 2^n$.

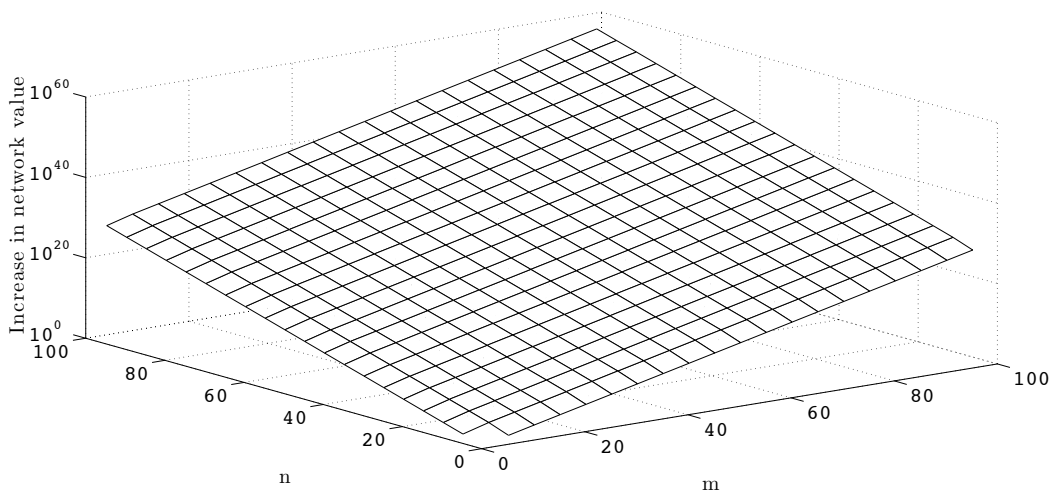


Figure 9: Increase in network value after interconnection or merging with Reed's law as valuation (logarithmic z-axis)

Sarnoff's law was not illustrated here, since this law does not provide any gain in network value due to interconnection or merging.

4 Findings from social networking services

This chapter contains findings from SNS, explanations on how the results were obtained and discussions of each topic. When data were obtained with a comprehensive method, the method is presented in the corresponding chapter. This applies to chapter 4.1 and 4.3. These chapters look at the relationship between content consumption and Zipf's law. A brief presentation of the specific SNS examined here is also given.

In chapter 4.5, 4.7 and 4.11 a Wikipedia article was used to find relevant information from SNS [17]. This article contains a list of the most common active SNS today. The SNS on this list were visited to obtain data about number of members, number of connections, average content created and estimated value. These statistics were retrieved on 4.13.2011 either through website information, request forms or emails. 203 SNS were visited and additional requests were sent to 57 SNS. A complete list of SNS contacted is listed in appendix A.

When a best-fit formula is given, it was calculated using IBM SPSS Statistics 19. Coefficient of Determination (R^2) values are also given when applicable. R^2 is the ratio of the explained variance (variance of the regression model) and the total variance (variance of actual data). R^2 is defined as:

$$R^2 = \frac{SS_{reg}}{SS_{tot}} = 1 - \frac{SS_{err}}{SS_{tot}} \quad (7)$$

Where

$SS_{tot} = \sum_{i=1}^n (y_i - \bar{y})^2$ is the total sum of squares

$SS_{reg} = \sum_{i=1}^n (f_i - \bar{y})^2$ is the explained sum of squares

$SS_{err} = \sum_{i=1}^n (y_i - f_i)^2$ is the residual sum of squares

y_i are actual observations

$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ is the average value of the actual observations

f_i are estimated values by the regression model

Best-fit formulas and R^2 values are given in chapter 4.1, 4.3, 4.7 and 4.11.

A residual is the distance of a point from the curve. A residual is positive when the point is above the curve and negative when the point is below the curve. When residuals from regression analysis are given in this chapter, they are transformed to studentized residuals. This is done since studentized residuals have two useful properties compared

to non-studentized residuals [18]:

- They have zero mean and unit standard deviation. This makes it possible to determine how far an observation is away from the mean in terms of standard deviation units.
- Leverage is a term used when some observations affect the outcome of a regression model significantly. Studentized residuals compensate for the leverage effect. Therefore, it is easier to observe residual outliers regardless of the leverage of the observations (outliers are residuals that are extremely far away from the regression curve in terms of standard deviation units).

Studentized residuals are calculated with the following formula:

$$Stud.Res = \frac{\epsilon_i}{\sqrt{MSE(1 - H_{ii})}} \quad (8)$$

Where:

ϵ_i is residual at observation i

$$\text{Mean Square Error (MSE)} = \frac{1}{n - m - 1} \sum_{i=1}^n \epsilon_i^2$$

n is the number of observations

m is the number of parameters in the regression model

H_{ii} is the diagonal elements of a hat matrix defined as:

$$H = X(X^T X)^{-1} X^T \quad (9)$$

where

$$X = \begin{bmatrix} 1 & x_1 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n \end{bmatrix}$$

The results from chapter 4.2, 4.4 and 4.8 will be discussed against the following assumptions for non-linear regression [19]:

1. Plausibility: the regression model is scientifically plausible.

2. Normality: the variability of values around the curve follows a Gaussian distribution.
3. Homoscedasticity: the response variables all have the same variance.
4. Accuracy: the model assumes that you know the independent variable(s) exactly.
5. Independence: the errors are independent of each other.

The results from 4.11 are not discussed against these assumptions, as the regression analysis in this chapter were based on few observations.

4.1 Zipf's law and Twitter

Andrew Odlyzko and Ben Tilly presumed that popularity in a network follows Zipf's law. The purpose of this chapter and chapter 4.3 was to see if such an approximation of popularity is valid in SNS.

Twitter (launched in 2006) is a free of charge social networking site with 175 million register users as of 2.15.2011 [20]. Members of the network can express their opinions and thoughts through text-based posts called "tweets". A member of the network can choose to follow any other Twitter member to receive their updates. Figure 10 shows the twitter profile of Bob Metcalfe, the inventor of Metcalfe's law. To the left in the figure, you see his most recent "tweets". Some facts about him, for example the number of members following Bob Metcalfe, is shown on the right side of the picture.

twitter Search Home Profile Messages Who To Follow

Bob Metcalfe
@BobMetcalfe UT: 30.286502,-97.742081
UTexas-Austin Professor of Innovation and Murchison Fellow of Free Enterprise; Polaris Partner; MIT Trustee. Past: Harvard, PARC, Stanford, 3Com, InfoWorld.

Follow

Timeline Favorites Following Followers Lists

BobMetcalfe Bob Metcalfe
 Today visiting CEO Melissa Simpler and team in Austin at <http://Affinegy.com>, which has installed 5+ million Digital Lifestyle Networks.
 2 hours ago

BobMetcalfe Bob Metcalfe
 Pitched <http://TuringScholars.org> about UTexas startup course. They offered pizza; Josh Baer handed out T-shirts - coins of the realm.
 3 hours ago

BobMetcalfe Bob Metcalfe
 Hearing "murdered civilians" in Afghanistan, wonder about choice of words. Most killing there is by people not wearing uniforms, right?
 13 hours ago

About @BobMetcalfe

4,216 Tweets 53 Following 6,161 Followers 547 Listed

Following 53

Similar to @BobMetcalfe · view all

SarahPalinUSA · Follow Sarah Palin

InghramAngle · Follow Laura Ingraham

bfeld · Follow Brad Feld

GarySinise · Follow Gary Sinise

About · Help · Blog · Mobile · Status · Jobs · Terms · Privacy · Shortcuts
 Advertisers · Businesses · Media · Developers · Resources · © 2011 Twitter

Figure 10: Bob Metcalfe's profile on Twitter

An Internet page keeps track of the 10 020 most popular Twitter users [21]. That is, the users with the most followers. The python script given at the next page was used to retrieve the statistics from the site.

```

import urllib2
import re

# Web page providing statistics
url_base = 'http://twittercounter.com/pages/100/'

# HTML string containing relevant data
reg_exp = '[0-9,]+</span> followers</div>'

# Create an empty array
followers = []

# Iterate through the 10020 most popular Twitter users
for i in range(0,10020,20):
    print str(float(i)/100)+'% complete'
    # Get the HTML file with the users with rank [i,i+19]
    html_content = urllib2.urlopen(url_base+str(i)).read()

    # Use regular expression to find right lines in the HTML file
    temp = re.findall(reg_exp, html_content);

    # Iterate through the relevant HTML files to find number of followers
    for j in range(0,len(temp)):
        if j != 0:
            temp[j] = temp[j].replace(',','')
            temp[j] = temp[j].replace('</span> followers</div>','')
            if int(temp[j]) != 0:
                followers.append(int(temp[j]))

# Avoid any inconsistencies by sorting the data
followers.sort(reverse=True)

# Writing the results to a text file
myfile = file("twitter.txt", 'w')
print >> myfile, followers
myfile.close()

```

The resulting data from the script were retrieved on 4.26.2011. Nine out of the 10 020 entries were obviously wrong as the number of followers was either 0 or out of order. This lead to a data basis of 10 011. Figure 17 shows the result where users are sorted descending by popularity.

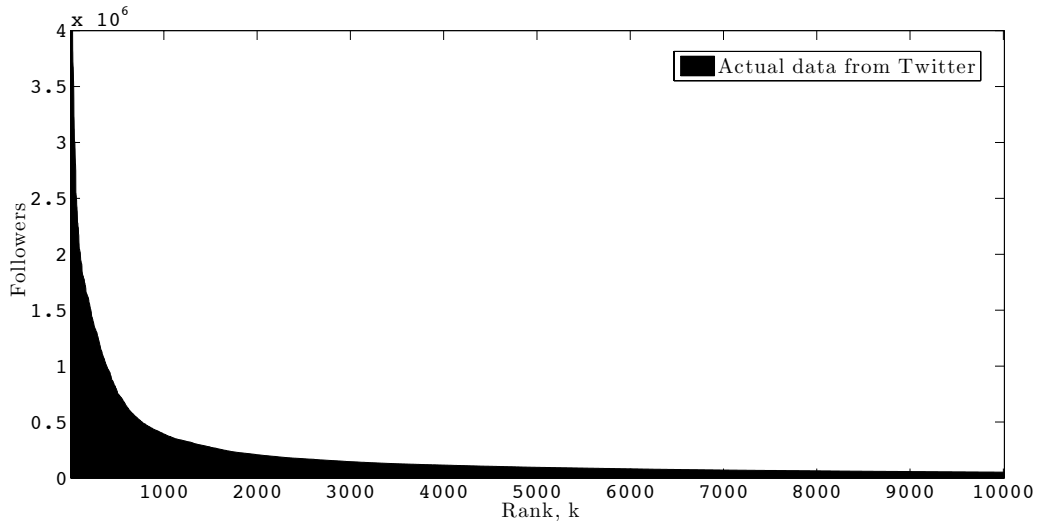


Figure 11: Number of followers for the 10 011 most popular twitter users

It is obvious that the data follows some sort of a long tail distribution. This implies that a power function may be suitable. Consequently Zipf’s law could also be a good fit, since this is a special type of a power function. To test the plausibility of different functions, best-fit formulas and corresponding R^2 values are calculated. These are given in table 4:

Fit type	Best-fit formula	R^2
Power	$1 \times 10^8 \times k^{-0.837}$	0.98953
Exponential	$446872 \times e^{-3 \times 10^{-4}k}$	0.78575
Logarithmic	$-4 \times 10^5 \times \ln(k) + 3 \times 10^6$	0.67679
Quadratic	$0.0256 \times k^2 - 329.72 \times k + 1 \times 10^6$	0.42312
Linear	$73.619 \times k + 583084$	0.23421

Table 4: Regression fits for data from Twitter, sorted by R^2

A power function fits the data with a very high correlation coefficient, so it seems likely that Zipf’s law also is a good fit. To be able to compare the popularity of Twitter users with Zipfs law, the data is transformed to frequency:

$$f_k = \frac{n_k}{2147851407}$$

Where 2147851407 is the total number of followers for the most popular 10 011 users and n_k the number of subscribers for user k. It is further assumed that the frequencies are sorted descending by popularity (f_1 is the most popular user, f_{10011} the least popular).

The value of the exponent (s), that fits the data best, is unknown. To find the optimal value of s, we need to minimize the Residual Sum of Squares (SS_{err}) function:

$$\min_s SS_{err} = \sum_{k=1}^{10011} \left(f_k - \frac{1/k^s}{\sum_{n=1}^{10011} 1/n^s} \right)^2 \text{ subject to } s > 0 \quad (10)$$

The Levenberg-Marquardt algorithm gives the optimal solution for Least Sum of Squares (SS_{opt}):

$$SS_{opt} = 0.0000636271$$

when

$$s = 0.56$$

This leads to the following best-fit formula for Zipf's law:

$$f(k, 0.56, 10011) = \frac{1/k^{0.56}}{\sum_{n=1}^{10011} (1/n^{0.56})} = \frac{1/k^{0.56}}{129.1195} \quad (11)$$

Note that the value of the exponent differs from the exponent for the best-fit power function. This is because Zipf's law has one degree of freedom more than a regular power function of the form αx^β . In the latter case, both α and β have to be estimated. The only parameter estimated with Zipf's law is the value of the exponent, s. Consequently; a power function will always have higher R^2 value, since it has one extra variable to adjust to improve the accuracy of the regression model.

The mathematics behind the calculation of R^2 after fitting the data from Twitter with Zipf's law follows:

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{10011} (y_i - f_i)^2}{\sum_{i=1}^{10011} (y_i - \bar{y})^2} = 1 - \frac{6.3706 \times 10^{-5}}{4.1940 \times 10^{-4}} = 0.8481 \quad (12)$$

Where:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{10011} f_i = 9.9894 \times 10^{-5} \quad (13)$$

An R^2 value of 0.8481 means that 84.81% of the variation on f_k can be explained by the regression on k .

Figure 12 shows the frequency of the data plotted against a function following a Zipfian distribution with $s = 0.56$ for $k = [1, 10011]$.

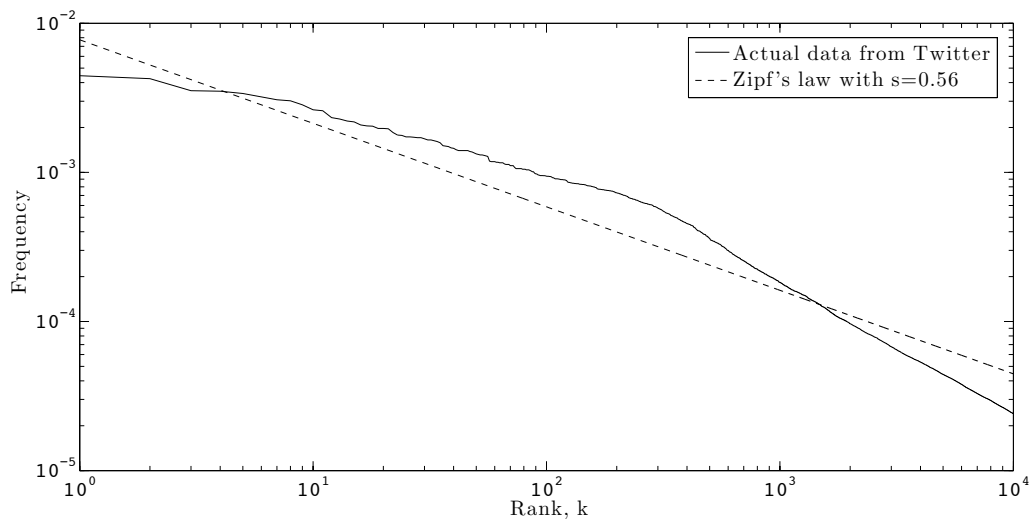


Figure 12: Fitting of Zipf's law and data from Twitter (logarithmic axes)

The same data are plotted in figure 13 with linear axes.

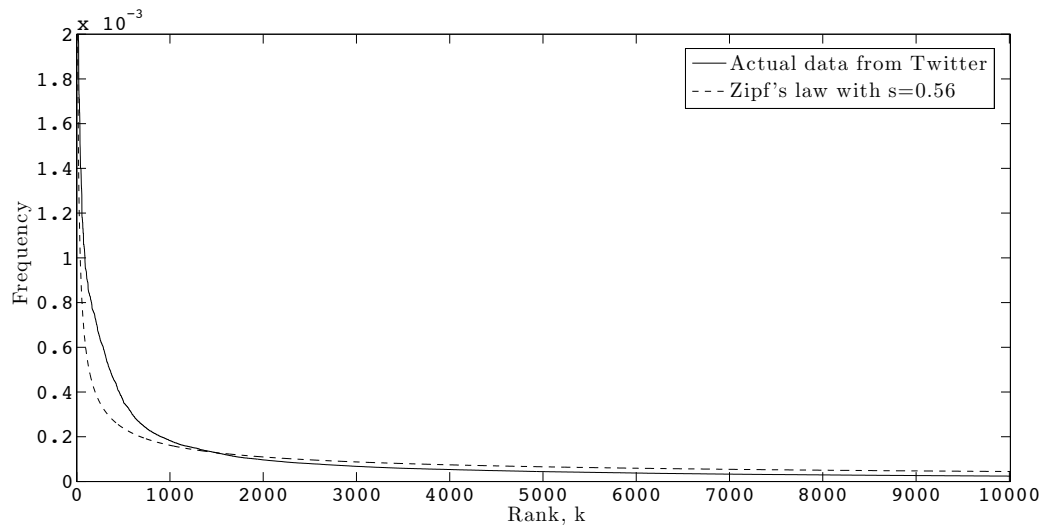


Figure 13: Fitting of Zipf's law and data from Twitter (linear axes)

The studentized residuals after fitting Zipf's law with Twitter are given in figure 14.

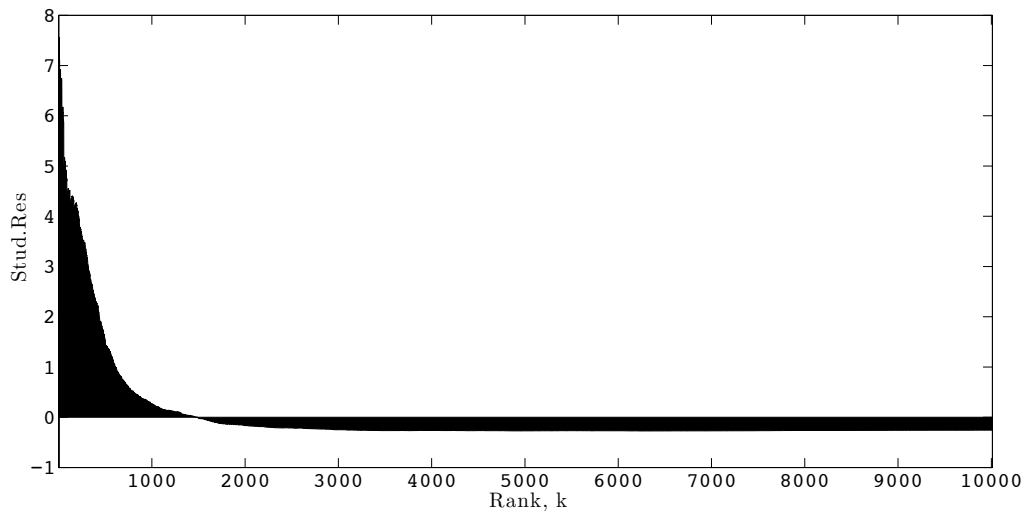


Figure 14: Studentized residuals after fitting data from Twitter with Zipf's law

Figure 15 displays a histogram of the studentized residuals.

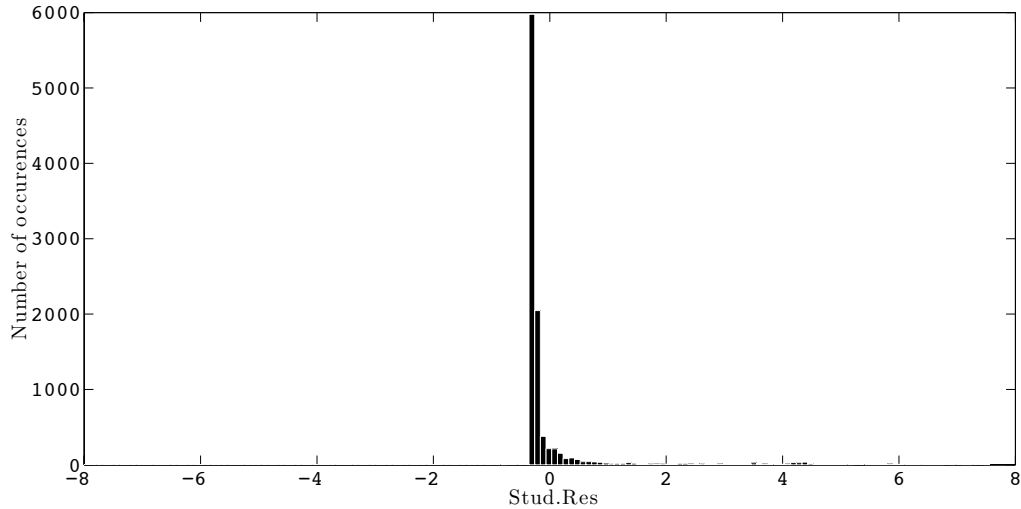


Figure 15: Occurrences of studentized residuals after fitting data from Twitter with Zipf's law

4.2 Discussion on Zipf's law and Twitter

Zipf's law was fitted with popularity of Twitter members since Twitter is a large and successful online community with a lot of relevant data available at [21]. Whether the results from the fitting violate the regression assumptions is discussed below:

1. Plausibility

Compared with the data from Twitter, Zipf's law with $s=0.56$ has an R^2 value of 0.8481. Even though this value does not indicate a very good fit, it does not mean that it is scientifically implausible to fit Zipf's law with the data from Twitter. Hence, the first regression assumption is met.

2. Normality

The occurrences of studentized residuals are not mirrored around origo, as it can be observed in figure 15. Given such a large sample size of 10011, the studentized residuals are expected to follow a Gaussian curve, given the normality assumption is fulfilled. This is not the case with the resulting studentized residuals from the fitting.

3. Homoscedasticity

The variance of studentized residuals in figure 14 is very high in the interval $[0,1000]$

compared to the rest of the interval [1000,10011]. At the most extreme, the standardized residuals are as high as seven times the standard deviation unit. This implies that the homoscedasticity property is violated.

4. Accuracy

The independent value, k , is known exactly, so this property is not violated.

5. Independence

There is a systematic pattern in the studentized residuals in figure 14. This implies that the studentized residuals are not independent of each other. This last property is therefore also violated.

As we have seen, several regression assumptions are violated. This implies that the data from Twitter cannot be fitted accurately with Zipf's law, at least not for the whole interval examined. Even though Zipf's law turned out to be an imprecise describer of the data from Twitter, a pure power function (αx^β) may be appropriate as its R^2 value indicated a very good fit. However, no further analysis where performed to conform this, as this was out of scope for this study.

4.3 Zipf's law and Youtube

Youtube was founded by Steve Chen, Chad Hurley and Jawed Karimin in 2005 [22] and bought one year later by Google. With its presence, the site made worldwide video sharing possible for anyone with an Internet connection. The main feature of the site is the possibility to share, watch and comment videos. The videos on the site are either uploaded by individuals or by site partners. Today, Youtube is one of the world's most visited website, with huge amounts of videos available. Figure 16 illustrates a video playback on Youtube.

Bob Metcalfe - Internet Pioneer / Entrepreneur

ComputerHistory 122 videos



0:27:52 / 1:29:03 360p

6,599

Uploaded by ComputerHistory on Apr 1, 2009

[Recorded: March 10, 2009]


34 likes, 1 dislikes


Show more


Top Comments

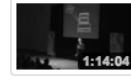
wow exelente video!!
ALONSO130 2 years ago 8


Suggestions


- 


Secret History of Silicon Valley
by ComputerHistory
70,569 views
- 

Technology & the Future of the Book
by ComputerHistory
4,617 views
- 

391 San Antonio - A Semiconductor Documentary
by ComputerHistory
12,693 views
- 

The Intel 80386 Business Case
by ComputerHistory
11,792 views
- 

Sara Blakely, Speaker, Entrepreneur & Founder ...
by BrooksInternational
38,620 views
- 

The History of Ethernet
by gidsr
25,142 views
- 

Donald Trump: Thought on Entrepreneurs
by luvvids007
308,472 views

Figure 16: Video playback on Youtube

James Zern, a software engineer at the company, revealed on 4.20.2011 that 99% of the views at Youtube come from 30% of the videos available [23]. This inequality indicates that video popularity is non-linear and that it might follow Zipf's law. To look further into this, a python script similar to the one written in chapter 4.1 was used to retrieve the statistics. The script, given at the next page, downloads the data from a Youtube page with the 160 all time most viewed videos [24].

```

import urllib2
import re

# Web page providing statistics
url_base = 'http://www.youtube.com/charts/videos_views?t=a&p='

# HTML string containing relevant data
reg_exp = '[0-9.]+ visninger '

# Create an empty array
youtube = []

# Iterate through the 160 most viewed Youtube videos
for i in range(1,9):
    print str(float(i)/0.08)+'% complete '

    # Get the HTML file with the users with rank [i,i+19]
    html_content = urllib2.urlopen(url_base+str(i)).read()

    # Use regular expression to find right lines in the HTML file
    temp = re.findall(reg_exp, html_content);

    # Iterate through the relevant HTML files to find number of views
    for j in range(0,len(temp)):
        temp[j] = temp[j].replace('.',',')
        temp[j] = temp[j].replace('<li class="last"><strong >',',')
        temp[j] = temp[j].replace('visninger ',',')
        youtube.append(int(temp[j]))

# Writing the results to a text file
myfile = file("youtube.txt", 'w')
print >> myfile, youtube
myfile.close()

```

The resulting data after running the script are illustrated in figure 17. The data were retrieved on 4.27.2011.

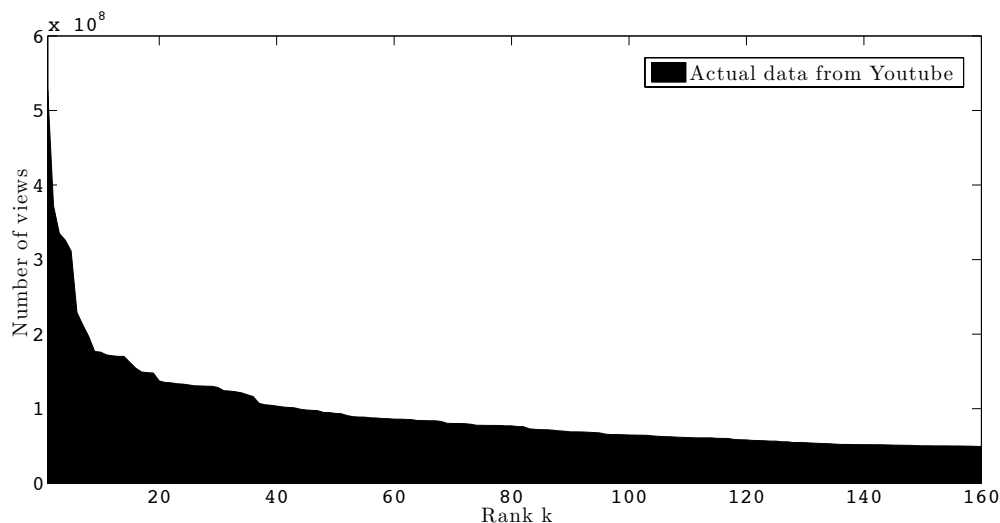


Figure 17: Number of views for the most popular Youtube videos

Like in section 4.1, different types of fit for the data are compared. The fit types and corresponding R^2 values are given in table 5.

Fit type	Best-fit formula	R^2
Power	$6 \times 10^8 \times k^{-0.485}$	0.9873
Exponential	$2 \times 10^8 e^{-0.009 \times k}$	0.85145
Logarithmic	$-6 \times 10^7 \times \ln(k) + 4 \times 10^8$	0.89573
Quadratic	$14481 \times k^2 - 3 \times 10^6 k + 2 \times 10^8$	0.7197
Linear	$-1 \times 10^6 k + 2 \times 10^8$	0.53176

Table 5: Several regression fits for data from youtube, sorted by R^2

A power function fits the data with a very high correlation coefficient, so it seems likely that the data can be fitted accurately with Zipf's law. To see how the data from Twitter follows a Zipfian distribution, frequency of each Youtube video is calculated as:

$$f_k = \frac{n_k}{15107824000}$$

Where 15 107 824 000 is the total number of views for the 160 most popular videos and n_k is the number of views for video of popularity rank k.

We need to solve an optimization problem similar as in chapter 4.1 to find the optimal value of the exponent, s :

$$\min_s SS_{err} = \sum_{k=1}^{160} \left(f_k - \frac{1/k^s}{\sum_{n=1}^{160} 1/n^s} \right)^2 \text{ subject to } s > 0 \quad (14)$$

The optimal solution for SS_{opt} can be calculated with the Levenberg-Marquardt algorithm:

$$SS_{opt} = 0.0000401594$$

when

$$s = 0.45$$

This leads to the following best-fit function for Zipf's law:

$$f(k, 0.45, 160) = \frac{1/k^{0.45}}{\sum_{n=1}^{160} (1/n^{0.45})} = \frac{1/k^{0.45}}{28.4102} \quad (15)$$

The corresponding R^2 value is calculated as:

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{160} (y_i - f_i)^2}{\sum_{i=1}^{160} (y_i - \bar{y})^2} = 1 - \frac{4.0160 \times 10^{-5}}{0.0028} = 0.9859 \quad (16)$$

Where:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{160} f_i = 0.0063 \quad (17)$$

An R^2 value of 0.9859 means that 98.59% of the variation on f_k can be explained by the regression on k .

Figure 18 shows the frequency of the data plotted against a function following a Zipfian distribution with $s = 0.45$ for $k = [1, 160]$.

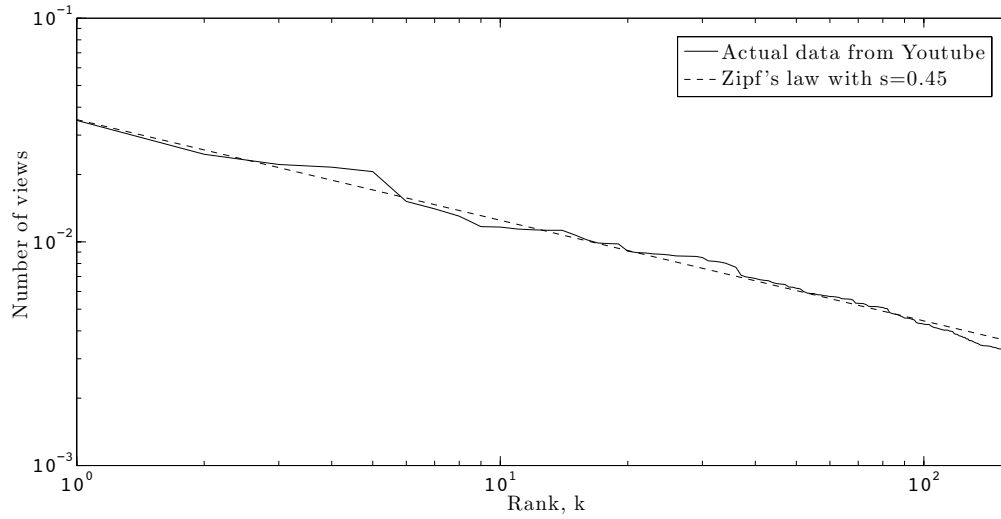


Figure 18: Fitting of Zipf's law and data from Youtube (logarithmic axes)

Figure 19 shows the same result, but with linear axes.

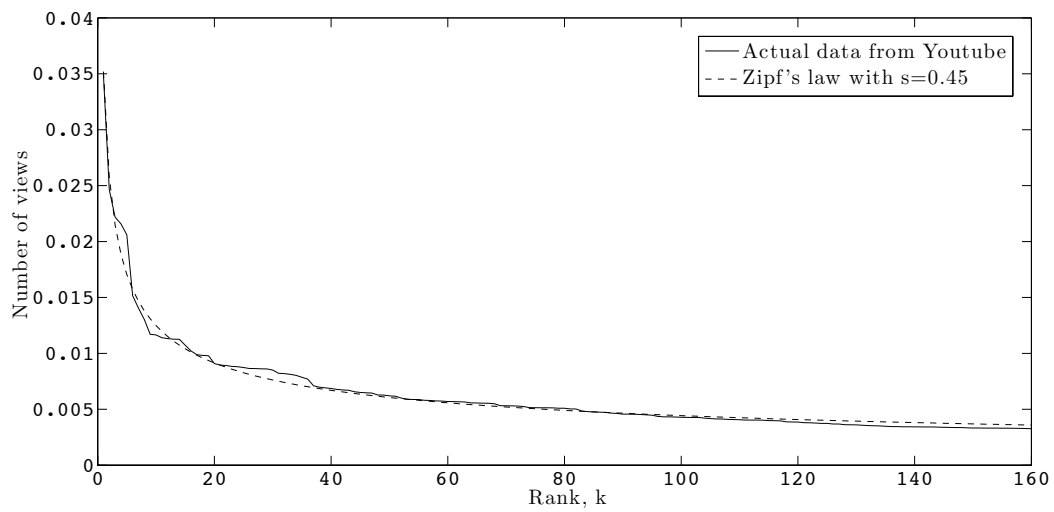


Figure 19: Fitting of Zipf's law and data from Youtube (linear axes)

The residuals from the fitting are given in figure 20.

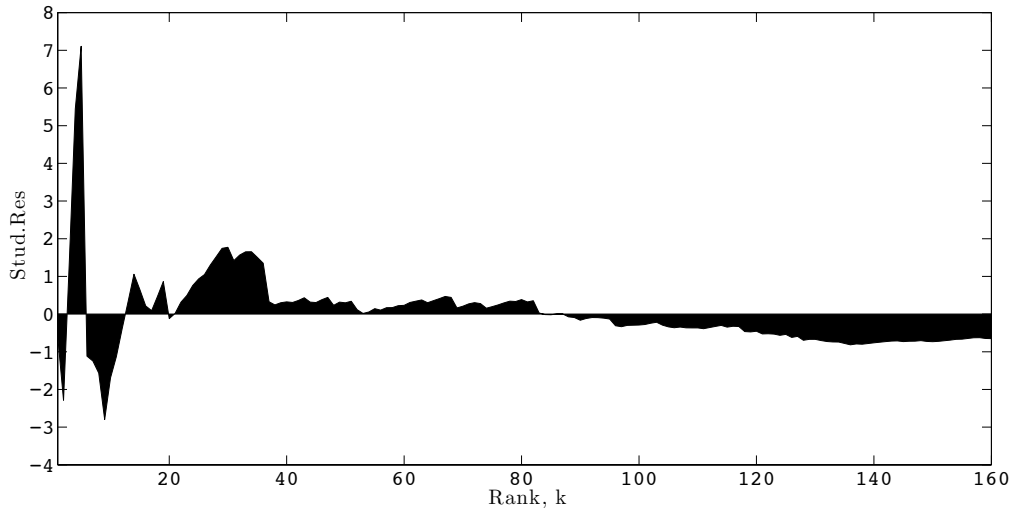


Figure 20: Studentized residuals after fitting data from Youtube with Zipf's law

Finally, a histogram of the studentized residuals is plotted in figure 21.

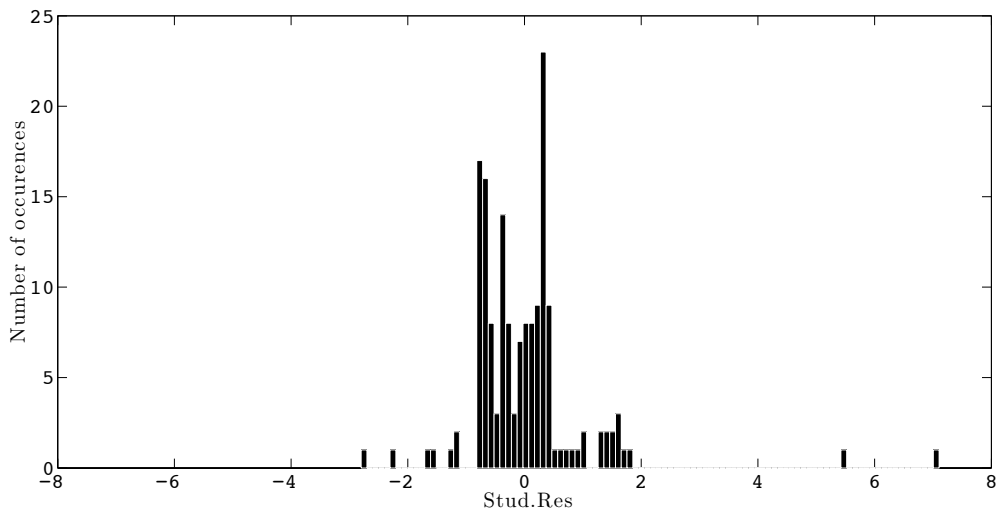


Figure 21: Occurrences of studentized residuals after fitting data from Youtube with Zipf's law

4.4 Discussion on Zipf's law and Youtube

Popularity of Youtube videos was compared with Zipf's law, since Youtube is the world's largest video-sharing site with data on popularity of videos available. The results are here discussed against the non-linear regression assumptions given in chapter 4 to discuss the trustworthiness of the fitting results:

1. Plausibility

The first regression assumption is met as Zipf's law with $s=0.45$ has a very high R^2 value (0.9859). This means that the regression line almost perfectly fits the data. It is therefore very likely that Zipf's law with $s=0.45$ fits to the data from YouTube.

2. Normality

The occurrences of studentized residuals are distributed approximately as a Gaussian distribution, as it can be observed in figure 21. The second requirement is therefore fulfilled.

3. Homoscedasticity

The variances of the studentized residuals are mostly equally divided in the interval, except for some outliers. These few studentized residuals were approximately seven times as high as the unit standard deviation. Overall, the homoscedasticity assumption was preserved with some exceptions.

4. Accuracy

The k values are known exactly, so this property is not violated.

5. Independence

The last regression requirement, independence, is not perfect, as the residuals are not completely randomly distributed. However, a systematic misfit, as observed with the fitting of Zipf's law with Twitter, is not registered here.

With regard to these points, the fitting of data from Youtube and Zipf's law seems to be a plausible estimate. Zipf's law had approximately the same R^2 value as a pure power function, despite having one extra degree of freedom.

4.5 Number of connections in a social network

The data used in this chapter were retrieved as described in the beginning of chapter 4. However, as it can be observed in table 6, limited data on average number of connections were available from SNS contacted. The entry marked with # in the table was retrieved by email.

Website	Number of members	Average number of connections
http://www.facebook.com/	500 000 000 [25]	130 [25]
http://twitter.com/	175 000 000 [26]	126 [27]
http://www.goodwizz.com/	75 000 #	20 #

Table 6: Social networking services with information about average number of connections and number of members

4.6 Discussion on number of connections in a social network

Several network laws proposed are based on potential connectivity. Therefore, the value grows $\Omega(n)$ with network size. Some observations of average number of connections are given in table 6. From this data it is obvious that interactions do not scale $\Omega(n)$ with network size, in contradiction to what several network laws propose. Even though the number of potential connections scales quadratically with network size in communication networks or even exponentially in GFNs, it does not seem to be a justifiable estimation of active connections in a network. Take Facebook as an example, a social networking site with approximately 500 million members in 2011 [25]. In this network, users are able to create various subgroups (through groups, events and community pages), which seem to fit Reed's definition of a GFN quite nice. Let us see how the number of potential subgroups differs from actual number of subgroups. The average user on Facebook is connected to 80 community pages, groups and events [25]. This leads to a maximum of: $\frac{500 \times 10^6 \times 80}{2} = 20$ billions actual subgroups, as at least two people are connected to a subgroup (by definition). If you even try to calculate the number of potential subgroups ($2^{500 \times 10^6}$), you probably either get infinite or an error telling the number is too large to be represented.

4.7 Relationship between content created and size of social networking services

This chapter looks at the relationship between network size and content created in SNS. When only average content created per month was available, 30.43684990 days per month was used for conversion between days and months. The results from a total of 15 SNS are given in table 7. SNS marked with * did only provide continuous day-to-day statistics. Data from SNS marked with # were retrieved by email.

Website	Members	Content created per day
http://www.facebook.com	500 000 000 [25]	1 478 471 003 [25]
http://www.twitter.com	175 000 000 [20]	140 000 000 [26]
http://www.badoo.com/	114 270 752 [28]	1 303 687 [28]
http://www.fotolog.com/ *	32 288 764 [29]	18 856 [29]
http://www.tumblr.com/	16 647 053 [30]	24 379 313 [30]
http://www.deviantart.com/	13 000 000 [31]	100 000 [31]
http://www.hyves.nl/ *	11 204 424 [32]	1 108 472 [32]
http://www.foursquare.com	8 000 000 [33]	2 500 000 [33]
http://www.meetup.com/	7 200 000 [34]	80 494 [34]
http://www.couchsurfing.org/ *	1 242 512 [35]	1 320 [35]
http://www.eproject-inc.com/ *	2 300 000 [36]	20 000 [36]
http://goodwizz.com	750 000 #	3 000 #
http://www.italki.com/	500 000 [37]	39 [37]
http://www.travellerspoint.com/ *	378 756 [38]	1 941 [38]
http://www.athlinks.com/	140 000 [39]	220 [39]

Table 7: Websites with information about content created and number of members

Table 8 shows various fit types for the data, sorted descending by the corresponding R^2 .

Fit type	Best-fit formula	R^2
Quadratic	$7 \times 10^{-9} \times n^2 - 0.5585 \times n + 5 \times 10^6$	0.99892
Linear	$2.7411 \times n - 5 \times 10^7$	0.90330
Power	$2 \times 10^{-8} \times n^{1.8555}$	0.79081
Exponential	$23444 \times e^{3 \times 10^{-8} \times n}$	0.47616
Logarithmic	$9 \times 10^7 \times \ln(n) - 1 \times 10^9$	0.28987

Table 8: Several regression fits for content created as a function of number of members

A comparison of quadratic, linear and power regression with actual data is given in figure 22.

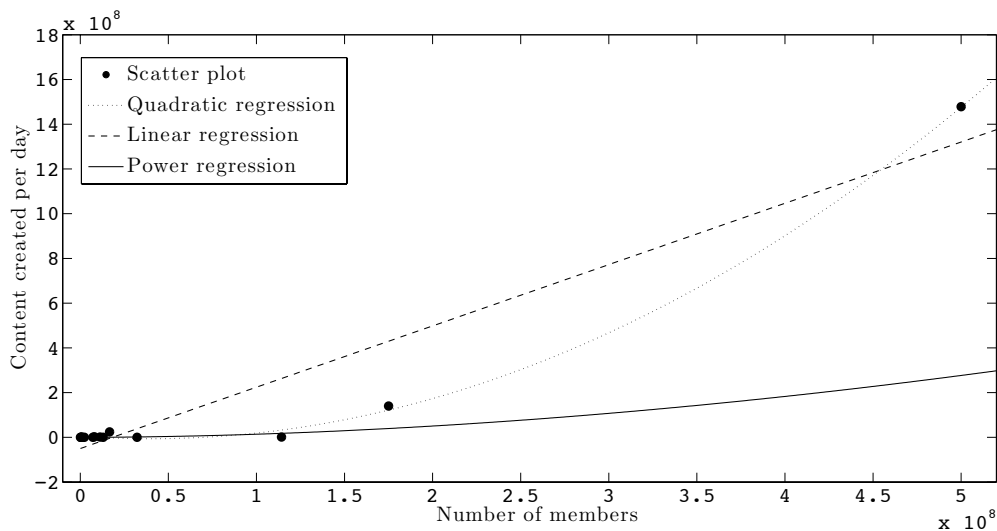


Figure 22: Comparison of quadratic, linear and power regression with actual data

A linear and quadratic fit seems both plausible, as both fit types has very high R^2 values. These two alternative fits are therefore studied in more detail. A closer look at the linear fit is given in figure 23.

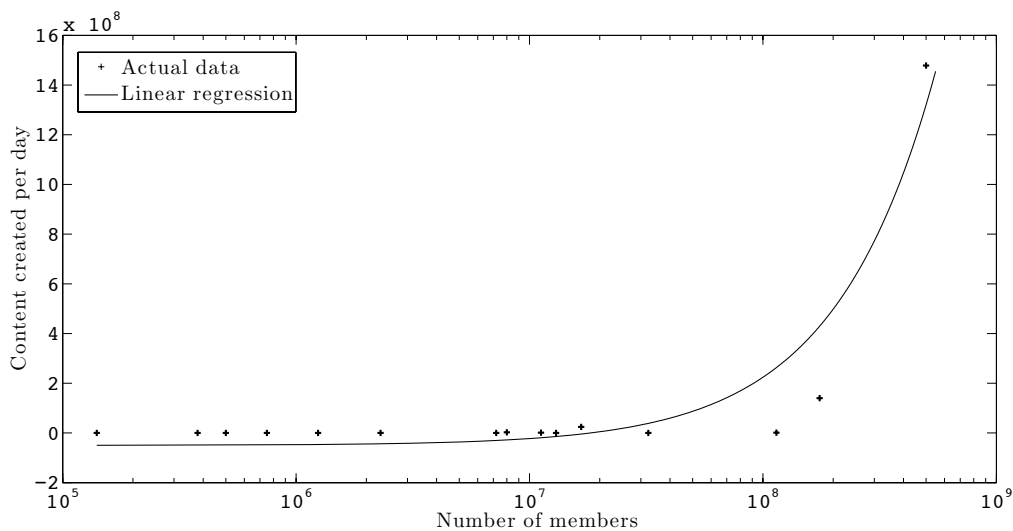


Figure 23: Linear regression (logarithmic x-axis)

The corresponding studentized residuals, from the linear fit, are given in figure 24.

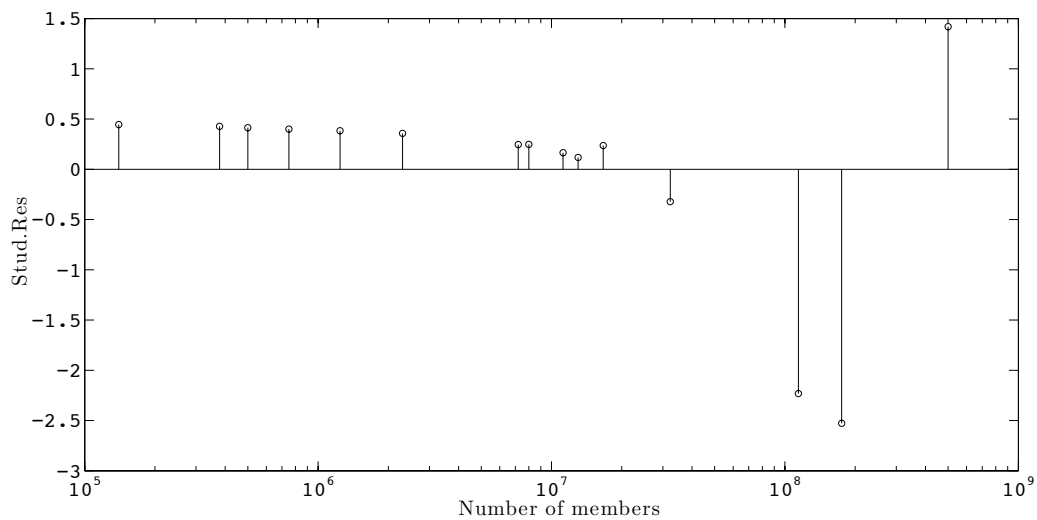


Figure 24: Studentized residuals from linear regression (logarithmic x-axis)

A histogram of the studentized residuals, with a linear fit, is given in figure 25.

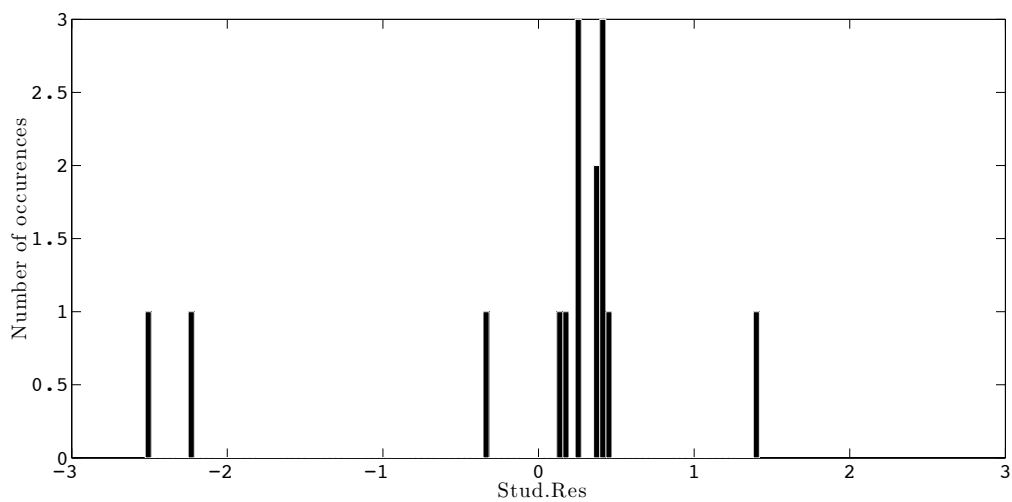


Figure 25: Occurrences of studentized residuals with linear regression

The quadratic fit is given in figure 26.

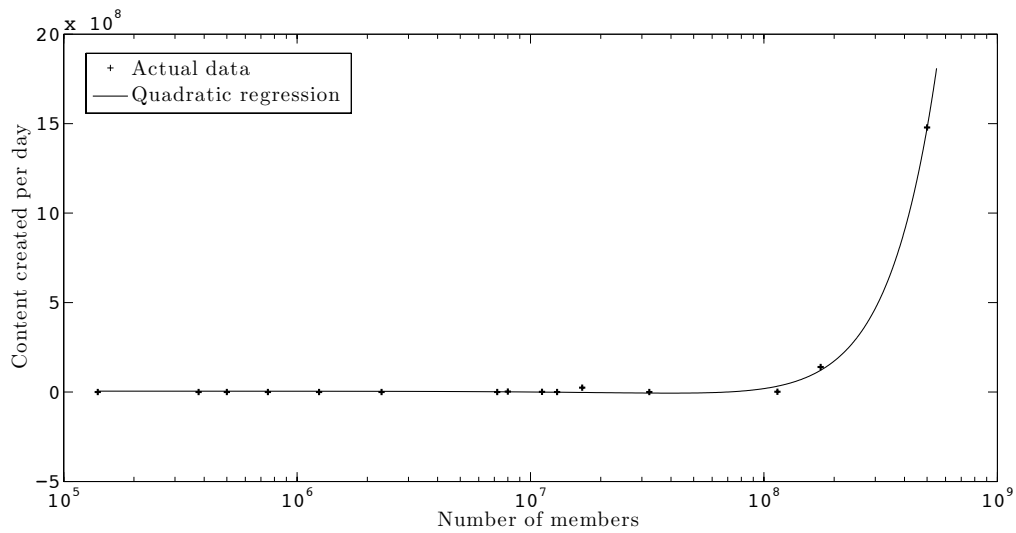


Figure 26: Quadratic regression (logarithmic x-axis)

The corresponding studentized residuals, with a quadratic fit, are given in figure 27.

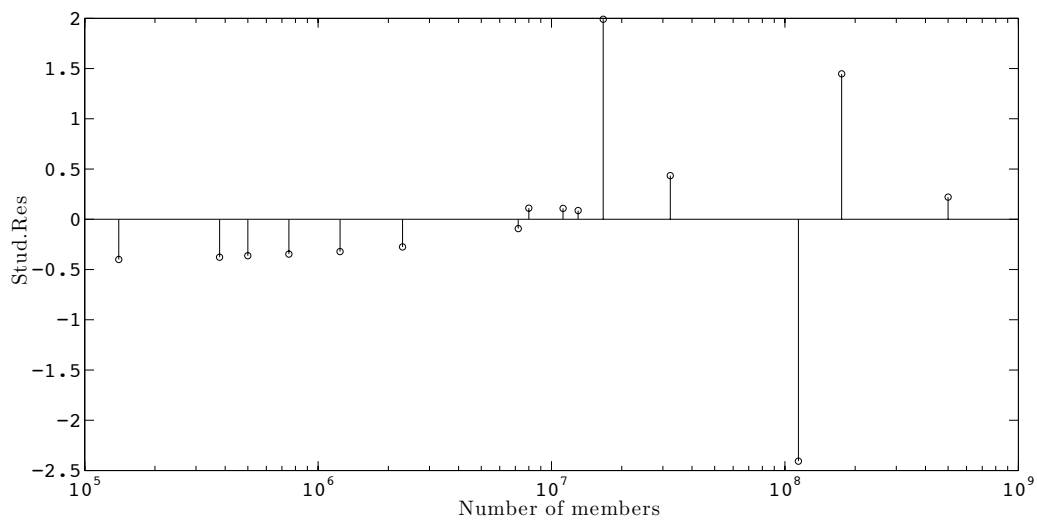


Figure 27: Studentized residuals from quadratic regression (logarithmic x-axis)

A histogram of the studentized residuals, with a quadratic fit, is given in figure 25.

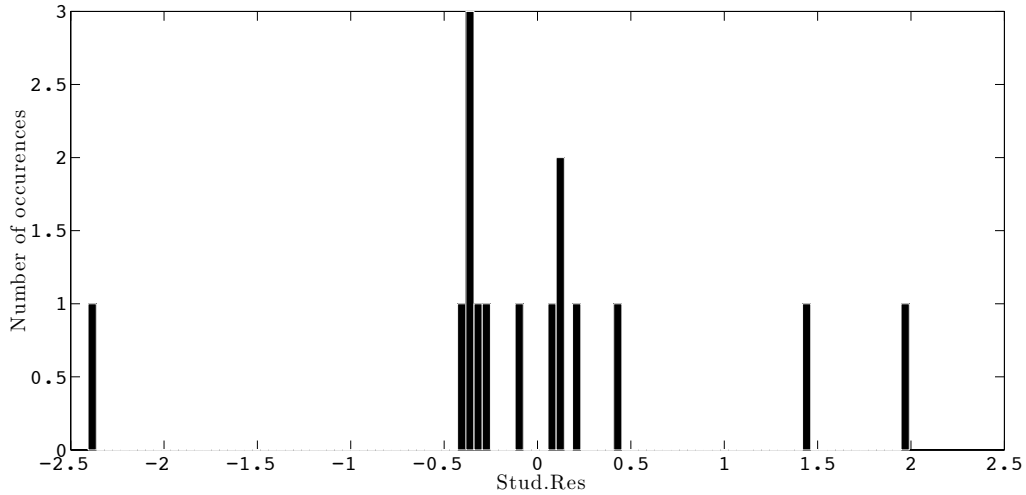


Figure 28: Occurrences of studentized residuals with quadratic regression

It is obvious that the quadratic fit has a lower or equal sum-of-squares value, since it has one extra parameter to adjust. However, an interesting question is whether the decrease in sum of squares is worth the loss in degrees of freedom with a quadratic function. An F-test can answer this question. To perform an F-test, two alternative hypotheses are needed:

H_0 : The best-fit for the data is a first order polynomial (straight line)

H_1 : The best-fit for the data is a second order polynomial (quadratic)

The relevant data for these calculations are given in table 9.

	Second order polynomial	First order polynomial	% Increase
Degrees of freedom	12	13	8.33 %
SS_{err}	2.188×10^{15}	1.958×10^{17}	8948.81 %

Table 9: Comparison of sum-of-squares and degrees of freedom

The F-ratio is defined as:

$$F = \frac{\% \text{ increase in sum - of - squares}}{\% \text{ increase in degrees of freedom}} \quad (18)$$

Applied to the data, the F-ratio is:

$$F = \frac{8948.81\%}{8.33\%} = 1062$$

An F ratio higher than 1.0 means that either:

- The more complicated model is correct (quadratic in this case).
- The simpler model is correct, but random errors made the complicated model to fit better than the simpler model.

The P-value describes the probability for the second case to be true. The confidence level for P is set to 0.05, corresponding to a 5% chance of rejecting the null hypothesis, given it was true. The null hypothesis will be rejected if $P < 0.05$.

The P-value, calculated with IBM SPSS Statistics 19 for the comparison of a first and second order polynomial, was < 0.0001 . Based on this, the H_0 hypothesis is rejected. Thus, a quadratic function fits the data significantly better than a linear function. It is therefore concluded that average productivity increases with network size for SNS studied.

4.8 Discussion on relationship between content created and size of networks

Some problems occurred during the work in this section. Few of the sites visited had available information about both network size and content created. In addition, network sizes from the sample data are not uniformly distributed as it can be observed in figure 22. Facebook, for example, has more members than the remaining 14 SNS summarized. Similarly, content created per day for Facebook exceeds content created for the rest of the 14 networks summarized. Consequently, there were few data points to fit in the interval $[0.5 \times 10^8, 5 \times 10^8]$ as it can be seen in figure 22.

The following section discusses the linear regression model against the regression assumptions listed at the beginning of chapter 4.

1. Plausibility

A linear fit would imply that productivity of a single network member is unaffected by network size, which seems to be a possible hypothesis. The linear fit has also a high R^2 value (0.9033) and is therefore plausible.

2. Normality

The occurrences of studentized residuals are centralized around 0, as it can be observed in figure 25. Since only 15 observations were made, it is expected that the studentized residuals do not follow a Gaussian curve perfectly. With this in mind, the normality requirement is fulfilled.

3. Homoscedasticity

The studentized residuals have approximately the same absolute value, except two outliers. These are approximately 2.5 times the size of the standard deviation unit as figure 24 shows.

4. Accuracy

When it comes to the accuracy requirement, it comes clear that some values of network size and content created are more precise than others, as some networks provided rounded numbers (as it can be seen in table 7). However, these numbers are provided by the SNS themselves and are therefore assumed to be good approximations.

5. Independence

The first 11 studentized residuals in figure 24 does not seem to be independent of each other, but again, only 15 data points were collected. It would be desirable to have more data to discuss further whether the studentized residuals were independent or not.

The following section discusses a quadratic regression model against the regression assumptions:

1. Plausibility

The quadratic model has an R^2 value of 0.99892. A quadratic fit would imply that productivity of a single network member increases with network size. This is also a plausible model and was therefore used as the alternative hypothesis. The first regression assumption is therefore met.

2. Normality

The occurrences of studentized residuals shown in figure 28 are approximately the same as with the linear regression. Therefore the same discussion applies as above.

3. Homoscedasticity

The studentized residuals in figure 27 are approximately of the same size with few exceptions, but again, only 15 data points were collected.

4. Accuracy

The accuracy requirement discussion is the same as with a linear function.

5. Independence

About half of the studentized residuals to the right in figure 27 seem to be independent of each other, while the first six are approximately the same. As mentioned earlier, it would be desirable to have more data to discuss further whether the studentized residuals were independent or not.

Both a linear and a quadratic fit, compared against the regression assumptions, seem to be plausible fitting functions. However, as the F-test gave an F-ratio is extremely high (1062) and significant at confidence level 0.05, it is concluded that a quadratic model is the best function to model content productivity as a function of network size.

4.9 What is important for members of online communities?

This chapter is simply a presentation of a study conducted by Petter B. Brandtzæg and Jan Heim. Their findings will later be discussed against existing network laws and results obtained in this study.

In "*User Loyalty and Online Communities: Why Members of Online Communities are not Faithful*", Petter B. Brandtzæg and Jan Heim studied why community-users stop using their social network [40]. The survey was conducted as an online survey within the following communities²:

Name	Number of members
Biip	280 000
Hamar-Ungdom *	190 000
Nettby*	320 000
Underskog	10 000

Table 10: Norwegian online communities in the study

The purpose of the survey was to reveal why online community members lack interest in an online community. The reasons for decreasing interest were grouped into nine specific and one cumulative category. Out of 200 responses, 257 reasons were given in total as

²Online communities marked with a * do no longer exist

some participants answered multiple reasons. The results from the online survey are given in table 11:

Reasons	Number of reported reasons (in %)
1. Lack of interesting people/friends attending	62 (24%)
2. Low quality content	59 (23%)
3. Low usability	45 (18%)
4. Harassment/bullying	24 (9%)
5. Time-consuming/isolating	16 (6%)
6. Low trust	15 (6%)
7. Over-commercialized	15 (6%)
8. Dissatisfaction with moderators	3 (1%)
9. Unspecified boring	3 (1%)
10. Other	15 (6%)
Total	257 (100%)

Table 11: Reasons why online community members stop using the social service or using it less

4.10 Discussion on what creates value in a social network

The results by Petter B. Brandtzæg and Jan Heim are here discussed against the networks laws presented earlier.

In tradition economy, utility is measured as the difference between maximum willingness to pay minus actual price paid. If we sum the utility for every user in a system, we get the total economic surplus or economic value. That is also the idea behind Beckstrom's law, but as pointed out earlier, the utility of network members is hard to measure in practice. So, how do we then measure value? In section 2, network externality was defined as *"networks where the utility of consumption is affected by the number of other users using the same or compatible products"*. Since the value of such networks directly depends on the number of users, active connections seem to be an indicator of network value. However, this number does not count for other categories that create value in SNS.

The study by Petter B. Brandtzæg and Jan Heim is an interesting contribution in the discussion that this chapter is devoted to. Their result reveals what is important for network members in some online communities in Norway. Network size seems to matter, as "lack of interesting people/friends attending" is the main reason why members value

the service less. However, this reason size alone does only stand for 24%. This result indicates that network laws based purely on network size do not count for the other 76% of what is important in Norwegian communities studied.

In the survey, "Low quality content" was a category with almost the same importance as "Lack of interesting people/friends attending". In SNS, the variety and quality of content is a key factor for success. Encouragement to produce content is therefore commonly observed in SNS. As an example of its importance, consider what the value of Twitter would be if the members stopped "tweeting". With this in mind, content productivity seems to be an important variable in the estimation of SNS value.

The remaining seven reasons in the study by Petter B. Brandtzæg and Jan Heim are harder to measure objectively. These categories are items where users typically have different requirements. For example, low trust, is probably a greater issue for people with higher technical competence. This makes these reasons harder to measure in practice. Still they count for the majority of the reasons with 53%. This result indicates that network laws purely based on network size are too simplistic.

4.11 Models for valuation of social networking services

This chapter presents alternative models for SNS value based on two variables: content created and network size. The data used in this chapter were retrieved as described in the beginning of chapter 4. Table 12 displays the data obtained. The following abbreviations are used in the table:

n_c = Current number of members

n_v = Number of members at valuation

c_c = Current content created per day

c_v = Content created per day at valuation

V = Market value

When data of content created at valuation are missing, c_v is estimated with the following formula:

$$c_v = \frac{n_v}{n_c} \times c_c$$

Social network	n_c	n_v	c_c	c_v	V
Facebook	500 000 000 [25]	-	14 785 000 000 [25]	-	\$ 41 000 000 000 [25]
Twitter	175 000 000 [26]	-	95 000 000 [26]	-	\$ 7 700 000 000 [41]
Fotolog	32 288 764 [29]	10 000 000 [42]	18856 [29]	5 840	\$ 90 000 000 [42]
Flixster	3 000 000 [43]	-	2 299 844 [44]	-	\$ 90 000 000 [45]
Badoo	114 270 752 [28]	-	142 609 [28]	-	\$ 300 000 000 [46]
Foursquare	8 000 000 [33]	1 800 000 [47]	2 500 000 [33]	562 500	\$ 95 000 000 [47]

Table 12: Size, content created and market value of social networking services

The data that are going to be modeled is illustrated in figure 29.

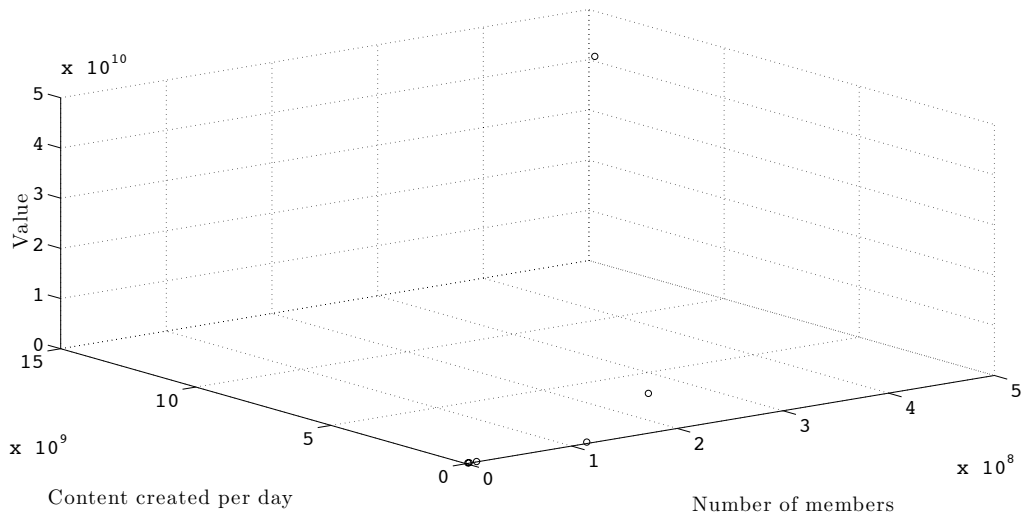


Figure 29: Scatter of actual value, average content created and network size

Three alternative fit formulas are tested against each other: a linear, a second-degree polynomial and a power fit. Figure 30, 31 and 32 display how content created correlates with market value, how number of members correlate with market value and how number of members correlate with content created, respectively. R^2 values are given for all fit types.

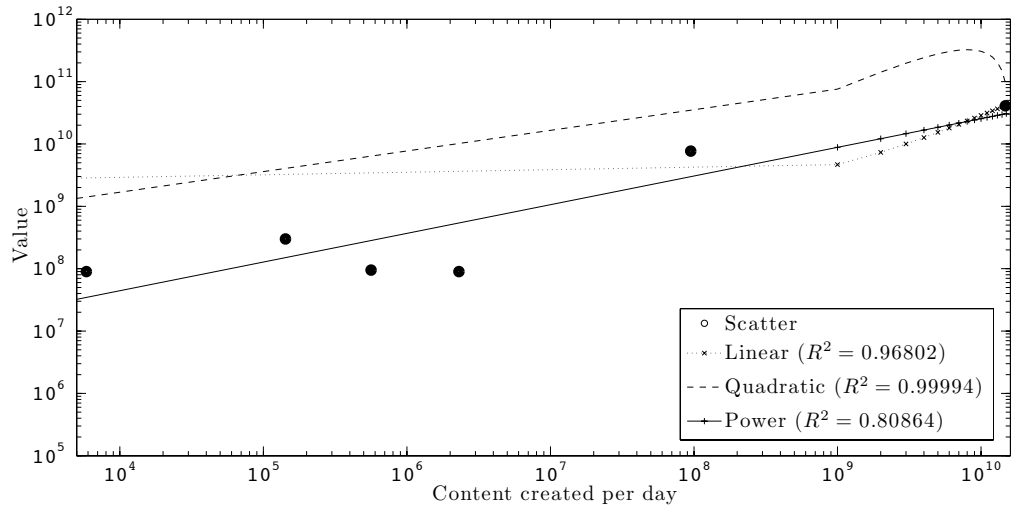


Figure 30: Correlation between content created and market value (logarithmic axes)

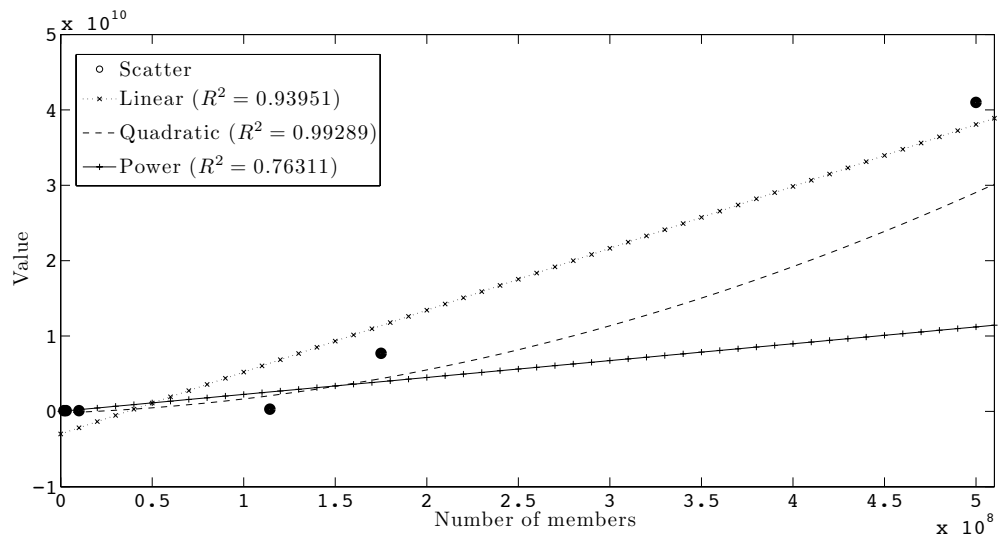


Figure 31: Correlation between number of members and market value

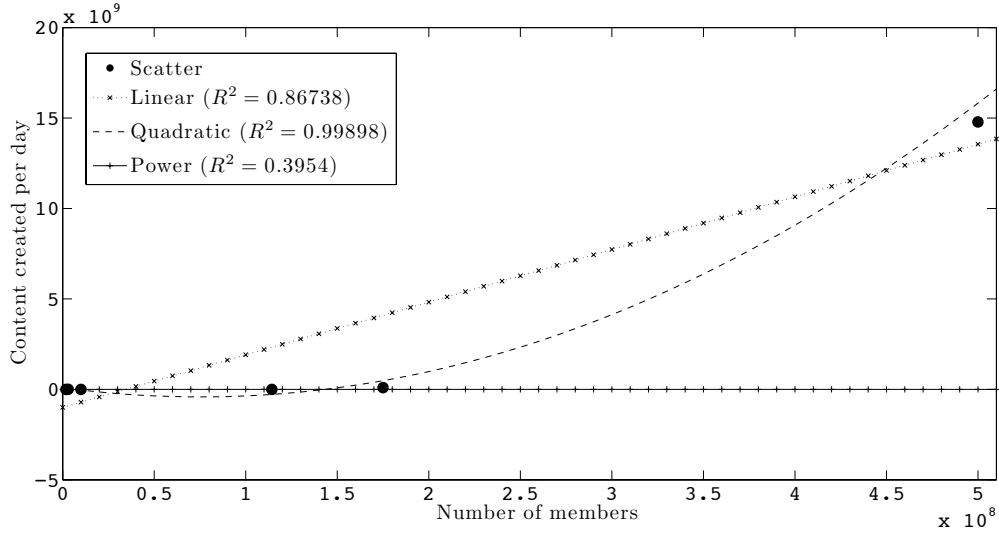


Figure 32: Correlation between number of members and content created

The first model estimated was a linear formula:

$$\bar{V}_{lrs}(n, c) = A + B_0 \times n + B_1 \times c \quad (19)$$

The unknown parameters A , B_0 and B_1 were calculated with Mathematica 8, using the function *FindFit*:

$$A = -4.38267 \times 10^8$$

$$B_0 = 33.8423$$

$$B_1 = 1.65914$$

This leads to the following best-fit linear response surface formula:

$$\bar{V}_{lrs}(n, c) = -4.38267 \times 10^8 + 33.8423 \times n + 1.65914 \times c \quad (20)$$

$\bar{V}_{lrs}(n, c)$ is plotted in figure 33.

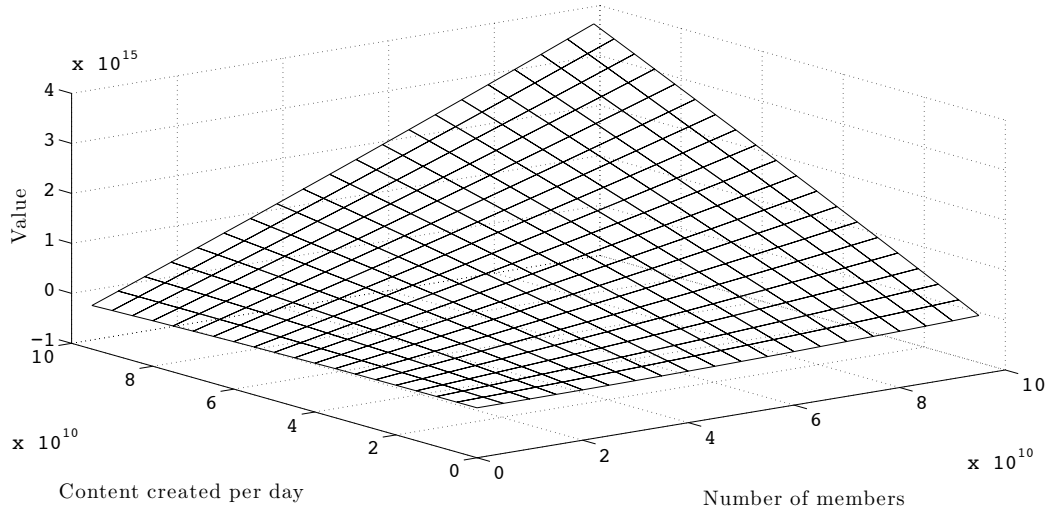


Figure 33: Response surface for estimated network value when \bar{V}_{lrs} is used as valuation

A closer look at the linear model applied to the data is given in table 13.

n	c	V	\bar{V}_{lrs}	V/\bar{V}_{lrs} ratio	Residual, ϵ
500000000	14785000000	41000000000	41013267900	0.999676497	13267900
175000000	95000000	7700000000	5641753800	1.364823825	-2058246200
10000000	5840	90000000	-99834310.62	-0.901493679	-189834310.6
3000000	2299843.782	90000000	-332924337.2	-0.270331694	-422924337.2
114270752	142609.4361	300000000	3429154679	0.087485117	3129154679
1800000	562500	95000000	-376417593.8	-0.252379277	-471417593.8
SS _{err}					1.44653 × 10 ¹⁹

Table 13: Residuals and accuracy of the linear regression model

The second alternative was a quadratic response surface of the form:

$$\bar{V}_{qrs}(n, c) = \beta_{00} + \beta_{10} \times n + \beta_{20} \times n^2 + \beta_{01} \times c + \beta_{02} \times c^2 + \beta_{11} \times n \times c \quad (21)$$

The calculated parameters were again calculated with Mathematica 8's *FindFit* function:

$$\beta_{00} = 9.85197 \times 10^7$$

$$\beta_{10} = -1.09755$$

$$\beta_{20} = 2.451 \times 10^{-8}$$

$$\beta_{01} = -3.69289$$

$$\beta_{02} = -1.49024 \times 10^{-8}$$

$$\beta_{11} = 4.52827 \times 10^{-7}$$

This leads to the following best-fit quadratic response surface:

$$\begin{aligned} \bar{V}_{grs}(n, c) = & 9.85197 \times 10^7 - 1.09755n + 2.451 \times 10^{-8} \times n^2 - 3.69289 \times c - \\ & 1.49024 \times 10^{-8} \times c^2 + 4.52827 \times 10^{-7} \times n \times c \end{aligned} \quad (22)$$

Figure 34 displays how value of SNS varies according to \bar{V}_{grs} .

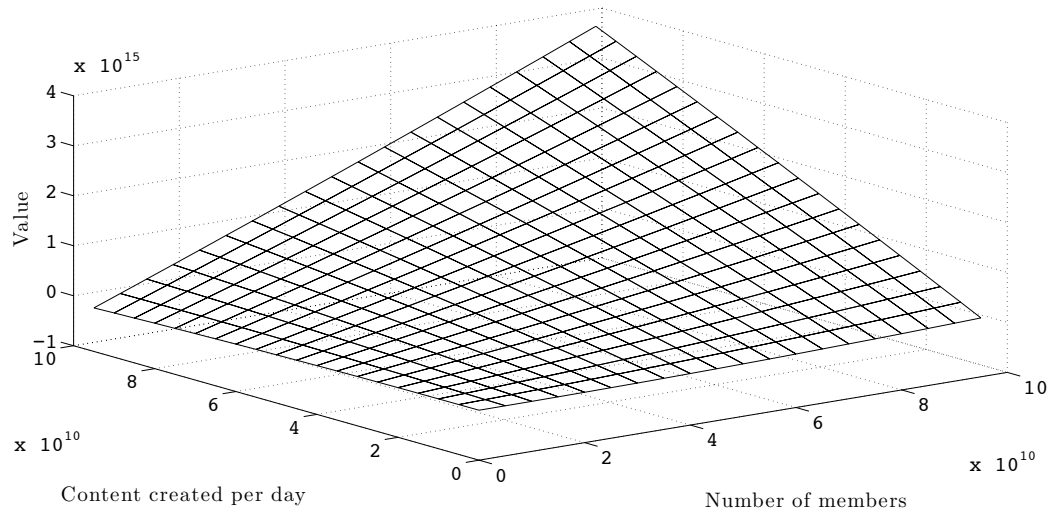


Figure 34: Response surface for estimated network value when \bar{V}_{grs} is used as valuation

Table 14 shows the residuals for the best-fit quadratic response surface.

n	c	V	\bar{V}_{grs}	V/\bar{V}_{grs} ratio	Residual, ϵ
500000000	14785000000	41000000000	40999998812	1.000000029	-1187.502663
175000000	95000000	7700000000	7699999727	1.000000035	-272.951536
10000000	5840	90000000	90000000	1.000000000	0.826648
3000000	2299843	90000000	89999999	1.000000011	-0.675020
114270752	142609	300000000	299999874	1.000000420	-125.397824
1800000	562500	95000000	94999999	1.000000011	-0.403340
SS_{err}					1498578

Table 14: Size, content created and market value of social networking services

Table 15 shows how network value varies with average content created per day and network size, according to \bar{V}_{grs} .

Content created ↓	Number of members				
	10 000	100 000	1 000 000	10 000 000	100 000 000
10 000	98,471,801.44	98,373,671.85	97,414,214.32	90,003,475.05	234,279,686.59
100 000	98,139,701.62	98,045,239.93	97,122,461.41	90,078,512.21	238,022,624.48
1 000 000	94,805,425.37	94,747,642.69	94,191,654.24	90,815,605.77	275,438,725.34
10 000 000	60,134,861.09	60,443,868.48	63,555,780.76	96,858,739.59	648,271,932.15
100 000 000	-419,350,960.33	-415,374,052.21	-375,583,132.63	24,509,899.19	4,243,823,821.60

Table 15: Estimated network value in USD with \bar{V}_{grs} for some common network sizes

The last alternative model was a power response surface of the form:

$$\bar{V}_{prs}(n, c) = \alpha \times n^{\beta_0} \times c^{\beta_1} \quad (23)$$

The best-fit power response surface was again calculated with Mathematica 8's *FindFit* function. The optimal variables were calculated to be:

$$\alpha = 14.1514$$

$$\beta_0 = 0.892437$$

$$\beta_1 = 0.167022$$

The optimal solution is given in equation 24:

$$\bar{V}_{prs}(n, c) = 14.1514 \times n^{0.892437} \times c^{0.167022} \quad (24)$$

The function is plotted in figure 35:

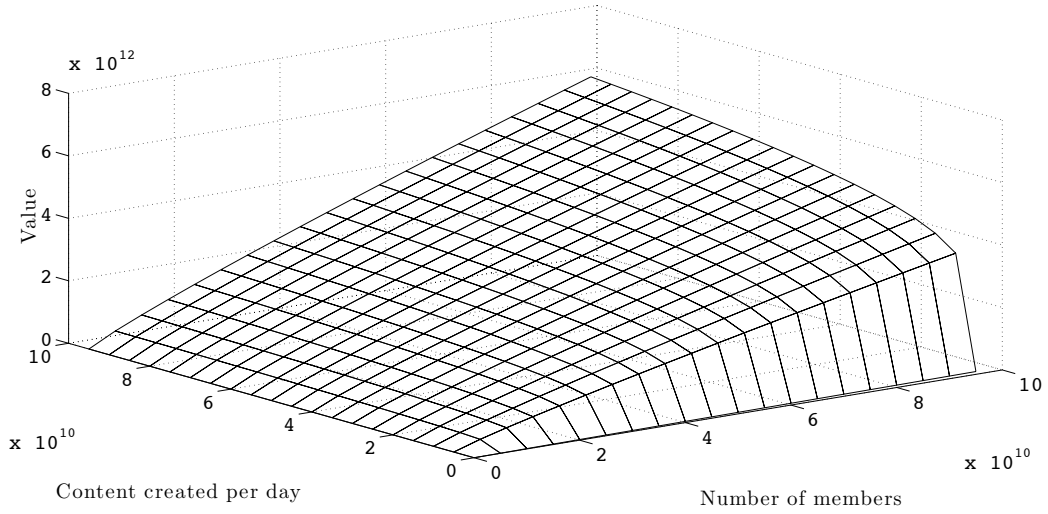


Figure 35: Response surface for estimated network value when $\bar{V}_{prs}(n, c)$ is used as valuation

Some calculation details about the best-fit power response surface are given in table 16.

n	c	V	\bar{V}	V/\bar{V}_{prs} ratio	Residual, ϵ
500000000	14785000000	41000000000	40988874992	1.000271415	-11125007.54
175000000	95000000	7700000000	6912687207	1.113893884	-787312793.4
10000000	5840	90000000	106397693.5	0.845882998	16397693.49
3000000	2299843.782	90000000	98574741.69	0.913012791	8574741.69
114270752	142609.4361	300000000	1595346308	0.188046945	1295346308
1800000	562500	95000000	49389314.69	1.923492978	-45610685.31
SS_{err}					2.30033×10^{18}

Table 16: $\bar{V}_{prs}(n, c)$ applied to data

Table 17 shows how network value varies with average content created per day and network size when $\bar{V}_{prs}(n, c)$ is applied as valuation formula.

Content created ↓	Number of members				
	10 000	100 000	1 000 000	10 000 000	100 000 000
10 000	1,127,374.08	4,296,037.31	16,370,729.95	62,383,256.86	237,721,271.42
100 000	2,228,784.32	8,493,135.33	32,364,436.09	123,329,804.86	469,967,736.40
1 000 000	4,406,238.94	16,790,670.70	63,983,507.56	243,819,279.94	929,111,946.69
10 000 000	8,711,000.61	33,194,646.22	126,493,451.92	482,023,314.11	1,836,826,110.87
100 000 000	17,221,383.75	65,624,807.78	250,073,714.13	952,945,457.83	3,631,349,455.36

Table 17: Estimated network value in USD with $\bar{V}_{prs}(n, c)$ for some common network sizes

It was earlier in this study estimated a relationship between content created and network size:

$$c(n) = 7 \times 10^{-9} \times n^2 - 0.5585 \times n + 5 \times 10^6$$

which in asymptotic terms is equal to:

$$c(n) = n^2 \tag{25}$$

Now, if we use equation 25 as an estimate for c in the formula $\bar{V}_{prs}(n, c)$, we get a model describing network value as a function of network size:

$$\bar{V}_{prs}(n, c) \approx n^{0.892437} \times c^{0.167022}$$

$$\bar{V}(n) = n^{0.892437} \times (n^2)^{0.167022} = n^{0.892437} \times n^{0.334044} = n^{1.226481}$$

In figure 36, $\bar{V}(n)$ is compared with other network laws only dependent on network size.

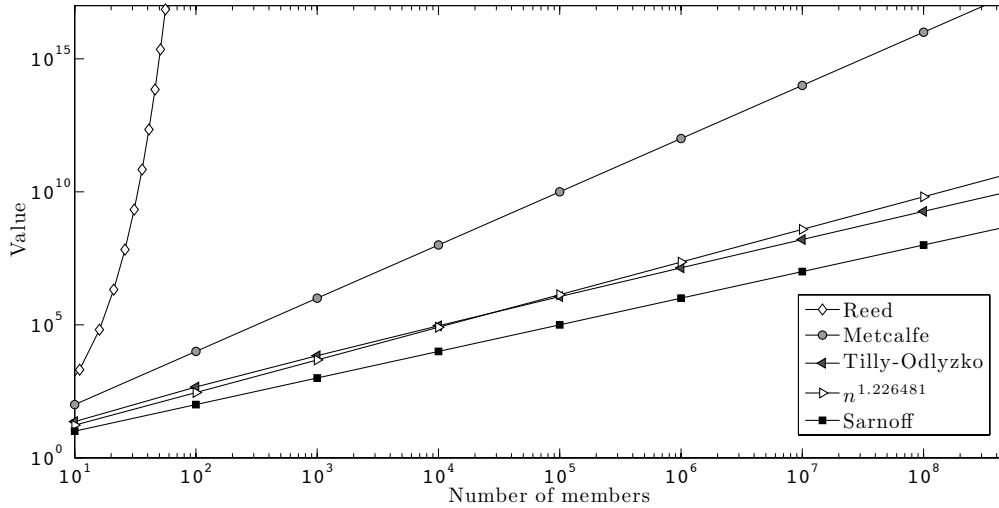


Figure 36: Comparison of $\bar{V}_{prs}(n)$ with existing network laws (logarithmic axes)

4.12 Discussion on models for valuation of social networking services

The purpose of this chapter was to give simple and precise models for valuation of SNS. The proposed models were based on actual market value, since this value represents what people actually are willing to pay. A benefit with this model is that it gives market value in a currency (United States Dollar (USD)). It is therefore clear what kind of value the model estimates, unlike several other network laws proposed.

Some problems were encountered during the work. The main problem was few observations available, despite extensive searches for data. Even though information requests were sent to 260 SNS, either passively or actively, only six SNS had the information needed. Since the models therefore were estimated on a limited data basis, it would be interesting to test the models against new data from SNS. This would clarify the model accuracy further. Another issue arose when network size, average content created and estimated value were not retrieved at the same date. Therefore a conversion was necessary (as described in section 4.11).

Three different models were proposed for valuation of SNS: a linear response surface, a quadratic response surface and a power response surface. The plausibility of these models are discussed in the next sections.

A linear response surface model had three estimated variables. This is what is positive

about the model, as it is simple to use. However, it is probably the only positive thing. Several negative values occurred during estimation, as table 13 shows. A linear surface model is therefore not an appropriate model for network valuation. Any further calculations were therefore not made with this formula. The main conclusion with the linear response surface model is that it is considered inappropriate for SNS valuation.

The next alternative model was a quadratic response surface. This model had six parameters fitted and is therefore vulnerable to overfitting, as only six data observations were made. Overfitting means that the model describes errors instead of actual relationships. This may occur when too many parameters are used compared to the number of observations. Is it likely that some of the data observations include errors? The observations used are from the SNS themselves or news articles. However, I suspect some of the numbers not to be accurate. As discussed earlier, some numbers are obviously rounded. In addition, the conversion described earlier has probably introduced some new errors. Overfitting is therefore likely to be a problem with this model. As expected, this model performed extremely good when it was tested against the observations used to find the best-fit function: the V/\bar{V}_{grs} ratio is extremely low, and consequently the SS_{err} value as well.

Table 15 illustrates network value for different values of average content created and network size according to \bar{V}_{grs} . An interesting observation in this table is that several values are negative. This is clearly an undesirable property. Another unwanted property is that estimated value sometimes decreases as content productivity or network size increases. The model is therefore considered inappropriate for SNS valuation. Any further calculations were therefore not conducted.

Finally, a power response surface was modeled. This model has also three estimated variables and is simple to use. This model does not have negative value estimations or decrease in value when productivity or network size increases. The model is therefore plausible. The V/\bar{V}_{prs} ratio given in table 16 shows the relative accuracy of the model. As it can be observed, the estimated value varies approximately between 18% and 192% of actual value. However, this is on the same data used for the estimation of the power response surface. Estimations with new data will probably be less precise off than this.

Table 17 displays estimated value when $\bar{V}_{prs}(n, c)$ is applied as valuation formula. According to this model, a doubling in network size increases network value approximately four times. A tenfold in average content created leads to approximately a doubling in network value.

A property of networks exhibiting a network effect is that a user number $n + 1$ provides

a network more value than user number n . This is because user number n gives utility to $n - 1$ members as they now are able to connect to entrant number n . Similarly; user number n is able to connect to $n - 1$ users. In contrast, users $n + 1$ brings more value to the network. This is because n users are able to connect to the new entrant. Similarly, user number $n + 1$ is able to benefit the n users in the network. As result of this, it is desirable that a valuation model grows $\Omega(n)$ when it is modeling networks exhibiting a network effect.

In chapter 4.11, the power response surface model, $\bar{V}_{prs}(n, c)$, was transformed to a function only dependent on network size. This way it was possible to discuss the power function proposed in this study against network effects. The proposed model exhibits network effects, as it grows faster than a linear function. The model grows much slower than both Reed and Metcalfe's law. This is a wanted property, as we have seen both Reed and Metcalfe's law lead to unrealistic network values. The model grows approximately as Tilly-Odlyzko's law, as it can be observed in figure 36. However, the model presented here is likely to be more accurate than Tilly-Odlyzko's law for SNS valuation, as the model presented in this study is based on actual observations of SNS value.

5 Summary

This section discusses the problems introduced in chapter 1.3.

- What generates value in networks?

Connectivity/network size has been the most common measure of network value in the network laws proposed. This decision is easy to justify since network effects are present in SNS.

No network laws proposed have used average content created as estimator of network value. As the study by Petter B. Brandtzæg and Jan Heim revealed, low quality content is an important reason why members loose interest in a social community. A social network with low content productivity implies that little new content is available. It also means low content diversity compared to a more productive network. Content productivity was therefore considered an important factor of network value and used as a variable in the SNS valuation models proposed in this study.

As mentioned earlier, several aspects other than those mentioned above affect individual utility/value in an online community. Some categories are harder to measure objectively, but what generates network value does not seem to be as monotonous as several network laws propose. The models proposed in this study deals with two variables, as these are both easy to measure and assumed important for network value.

- Is each network connection of equal value?

In SNS, people that create content are more likely to be of greater importance than people that only consume content. The 1% rule indicates that only a small proportion of the community create content. Those participants are likely to be of higher importance than a non-contributor. Given the 1% rule is valid, a scheme where each network connection is of equal value is unsuitable. Similarly, Zipf's law indicates that popularity of content/connections follows a Zipfian distribution. An equal value of each network connection does not fit this theory either. An equal network connection value is only justifiable when it comes to potential connectivity.

- Is it likely that one network law can accurately describe the value of all networks?

As several SNS have been introduced, it might have come clear to the reader that SNS differ very much. Twitter, for example, does only provide text-based updates. Facebook, in contrast, has great service diversity. It is therefore important

to remember that the models proposed only account for some of the variables that contribute to network value. Several other factors influence value, not taken into account in here. When it has been observed such a great diversity within social networks, an even greater diversity is expected if other network domains are considered. Thus, if precise estimations are needed, domain specific network laws are most appropriate.

This said, the purpose of the model is not to be a perfect estimator. This would require the model to be far too complex to have any practical purpose, both in terms of difficult variables to measure and difficulty of calculation.

- How do you test the accuracy of a network law?

Most of the network laws proposed are theoretical and have therefore to a limited extent been tested against real world network properties. Tilly and Odlyzko used real world cases of interconnection and merging to test the accuracy of their law. This may give an indicator of whether the law is plausible, but no specific benchmarks. The models presented in this study are empirical, and should therefore model the real world. The value estimated by the models in this study can be tested against known market value to give an exact performance benchmark.

6 Conclusions and further work

6.1 User behavior in social networking services

- The best-fit formula to model popularity of Twitter members as a Zipfian function was:

$$f(k, 0.56, 10011) = \frac{1/k^{0.56}}{129.1195}$$

The corresponding R^2 value was 0.8481. This implies that Zipf's law is a plausible model, but not an ideal fit. After further analysis of the studentized residuals it was concluded that popularity of Twitter users could not be fitted accurately with Zipf's law. At least not for the whole interval examined.

- The best-fit formula to model popularity of Youtube videos as a Zipfian function was:

$$f(k, 0.45, 160) = \frac{1/k^{0.45}}{28.4102}$$

The corresponding R^2 value was 0.9859. This implies that Zipf's law is a very good estimation of popularity of Youtube videos. The analysis of the studentized residuals supported this conclusion, as the residuals to a large extent were in accordance with the regression assumptions.

- A power function, but not necessarily Zipf's law, is probably the best option to model popularity of content in SNS.
- A person's number of connections in SNS does not scale $\Omega(n)$.
- The best-fit function to model content productivity (average content created per day) as a function of network size was:

$$c(n) = 7 \times 10^{-9} \times n^2 - 0.5585 \times n + 5 \times 10^6$$

This quadratic function was significantly better than the best-fit linear function. Thus, average productivity increases with network size for SNS studied.

6.2 Valuation of social networking services

- Giving precise value estimations of SNS are hard.
- A linear response surface model and a quadratic response surface model were found inappropriate for SNS valuation.
- The most appropriate model for valuation of SNS was the power response surface:

$$\bar{V}_{prs}(n, c) = 14.1514 \times n^{0.892437} \times c^{0.167022}$$

This model uses number of members and average content created as variables.

- As a function of network size, the proposed model grew $n^{1.226481}$ in asymptotic terms - approximately as Tilly-Odlyzko's law.

6.3 Further work

- Test Zipf's law against popularity of content in more SNS.
- Gather more data (used during the estimation of a power response surface model) to either:
 - Estimate the variables in a power response surface more accurately.
 - Test the proposed valuation model further.

References

- [1] Os Shy (2001), *The Economics of Network Industries*, Cambridge University Press
- [2] Andrew Odlyzko, Benjamin Tilly (2005), *A refutation of Metcalfe's Law and a better estimate for the value of networks and network interconnections*. Available at: <http://www.dtc.umn.edu/~odlyzko/doc/metcalfe.ps>. Accessed 2.14.2011
- [3] Robert M. Metcalfe (2006), *Metcalfe's Law Recurses Down the Long Tail of Social Networks*. Available at: <http://vc mike.wordpress.com/2006/08/18/metcalfe-social-networks/>. Accessed 1.27.2011
- [4] Stephen Shankland (2005), *Researchers: Metcalfe's Law overshoots the mark* <http://www.zdnet.com/news/researchers-metcalfes-law-overshoots-the-mark/141783>. Accessed 1.27.2011
- [5] Bob Briscoe, Andrew Odlyzko, Benjamin Tilly (2006), *Metcalfe's Law is Wrong*. Available at <http://spectrum.ieee.org/computing/networks/metcalfes-law-is-wrong>. Accessed 1.27.2011
- [6] CTIA (2009), *Background on CTIA's Semi-Annual Wireless Industry Survey* Available at: http://files.ctia.org/pdf/CTIA_Survey_Midyear_2009_Graphics.pdf. Accessed 9.2.2011
- [7] David P. Reed (1999), *That Sneaky Exponential - Beyond Metcalfe's Law to the Power of Community Building*. Available at: <http://www.ate.co.nz/networking/reedslaw.htm>. Accessed 1.27.2011
- [8] Rod Beckstrom (2009), *A New Model for Network Valuation*, National Cybersecurity Center. Available at: http://www.beckstrom.com/The_Economics_of_Networks. Accessed 1.27.2011
- [9] <http://blog.postmaster.gr/2009/04/08/beckstroms-law-fail/>. Accessed 2.11.2011
- [10] <http://tech.slashdot.org/comments.pl?sid=1190037&cid=27488055>. Accessed 2.11.2011
- [11] George Kingsley Zipf (1932), *Selected Studies of the Principle of Relative Frequency in Language*, Cambridge, MA.: Harvard University Press.
- [12] George Kingsley Zipf (1949), *Human Behavior and the Principle of Least Effort*. Cambridge, MA Addison-Wesley

- [13] Clay Shirky, *Power Laws, Weblogs, and Inequality*. Available at: http://www.shirky.com/writings/powerlaw_weblog.html. Accessed 2.13.2011
- [14] Lada A. Adamic and Bernardo A. Huberman (2002), *Zipf's law and the Internet*. Available at: <http://access.cs.sci.ku.ac.th/~usa/418591/2010-1/practice/sukumal/Zipf-internet.pdf>.
- [15] <http://www.antseyeview.com/90-9-1-principle/principle-in-action/>. Accessed 3.31.2011
- [16] Osamuyimen Stewart, David Lubensky and Juan M Huerta, *Crowdsourcing Participation Inequality : A SCOUT Model for the Enterprise Domain*
- [17] http://en.wikipedia.org/wiki/List_of_social_networking_websites. Accessed 2.21.2011
- [18] Rudolf J. Freund, William J. Wilson and Ping Sa, *Regression Analysis*. Second Edition. Academic Press
- [19] <http://www.graphpad.com/www/nonling3.htm>. Accessed 5.13.2011
- [20] <http://twitter.com/about>. Accessed 2.15.2011
- [21] <http://twittercounter.com/pages/100/>. Accessed 2.21.2011
- [22] http://www.usatoday.com/tech/news/2006-10-11-youtube-karim_x.htm. Accessed 4.28.2011
- [23] <http://www.telegraph.co.uk/technology/news/8464418/Almost-all-YouTube-views-come-from-just-30-of-films.html>. Accessed 4.28.2011
- [24] http://www.youtube.com/charts/videos_views?t=a. Accessed 4.27.2011
- [25] <http://www.facebook.com/press/info.php?statistics>. Accessed 2.15.2011
- [26] <http://techcrunch.com/2011/03/14/new-twitter-stats-140m-tweets-sent-per-day-460k-accounts-created-per-day/>. Accessed 4.12.2011
- [27] <http://www.guardian.co.uk/technology/blog/2009/jun/29/twitter-users-average-api-traffic>. Accessed 3.14.2011
- [28] <http://corp.badoo.com/company/>. Accessed 4.13.2011
- [29] <http://www.fotolog.com/>. Accessed 4.12.2011
- [30] <http://www.tumblr.com/about>. Accessed 4.12.2011
- [31] <http://about.deviantart.com/>. Accessed 4.12.2011

- [32] <http://www.hyves.nl/>. Accessed 4.12.2011
- [33] <https://foursquare.com/about>. Accessed 4.12.2011
- [34] <http://www.meetup.com/about/>. Accessed 4.12.2011
- [35] http://www.couchsurfing.org/mission_stats.html. Accessed 4.12.2011
- [36] <http://www.eproject-inc.com/>. Accessed 4.12.2011
- [37] <http://www.italki.com/static/about.htm>. Accessed 4.12.2011
- [38] <http://www.travellerspoint.com/>. Accessed 4.12.2011
- [39] <http://www.athlinks.com/>. Accessed 4.12.2011
- [40] Petter B. Brandtzæg and Jan Heim, *User Loyalty and Online Communities: Why Members of Online Communities are not Faithful*
- [41] <http://www.blogherald.com/2011/03/07/twitter-wooth-7-7-billion-probably-not>. Accessed 4.5.2011
- [42] <http://gawker.com/#!293778/confirmed/frances-hi+media-buys-fotolog-for-90-million>. Accessed 4.5.2011
- [43] <http://www.flixster.com/about>. Accessed 4.12.2011
- [44] <http://www.flixster.com/about/advertise>. Accessed 4.12.2011
- [45] <http://latimesblogs.latimes.com/technology/2011/03/warner-bros-could-buy-flixster-in-yet-another-social-media-move.html>. Accessed 5.4.2011
- [46] <http://eu.techcrunch.com/2008/01/21/uks-badoo-pulls-30m-for-russian-launch-ahead-of-a-home-push/>. Accessed 4.5.2011
- [47] <http://techcrunch.com/2010/06/29/foursquare-20-million/>. Accessed 4.5.2011
- [48] <http://www.ieice.org/~wtc2012/index.html>. Accessed 6.6.2011

A Request sent to various social networking services

A.1 Social networking services contacted

<http://www.anobii.com/>
<http://www.athlinks.com/>
<http://www.asmallworld.net/>
<http://www.audimated.com/>
<http://badoo.com/>
<http://www.bebo.com/>
<http://blauk.com/>
<http://www.cafemom.com/>
<http://www.care2.com/>
<http://www.couchsurfing.org/>
<http://www.cozycot.com/>
<http://www.cross.tv/>
<http://www.crunchyroll.com/>
<http://dailybooth.com/>
<http://www.deviantart.com/>
<http://www.disaboom.com/>
<http://www.experienceproject.com/>
<http://www.exploroo.com/>
<http://fledgewing.com/>
<http://www.flixster.com/>
<http://www.fotolog.com/>
<http://www.gamerdna.com/>
<http://www.gogoyoko.com/>
<http://www.goodwizz.com/>
<http://www.thehotlist.com/>
<http://identi.ca/>
<http://www.italki.com/>
<http://www.internations.org/>
<http://kaioo.com/>
<http://lafango.com/>
<http://www.lifeknot.com/>
<http://www.listography.com/>
<http://www.livejournal.com/>
<http://www.mocospace.com/>
<http://mog.com/>
<http://www.mouthshut.com/>
<http://mubi.com/>
<http://www.multiplay.co.uk/>
<http://muxlim.com/>
<http://netlog.com/>
<http://www.nexopia.com/>
<http://oneworldgroup.org/tv>
<http://www.opendiary.com/>
<http://www.playlist.com/>
<http://www.plurk.com/>
<http://raptr.com/>
<https://www.ravelry.com/>
<http://sciencestage.com/>
<http://www.sharethemusic.com/>
<http://www.stumbleupon.com/>
<http://zooppa.com/>
<http://www.teachstreet.com/>
<http://www.travellerspoint.com/>
<http://social.wakoopa.com/>
<http://www.webbiographies.com/>
<http://www.worldfriends.no/>
<https://www.xing.com/>

A.2 Email request

Date: 04.13.2011 12:55

Subject: Master thesis

To whom it may concern,

I am writing a master thesis at the Norwegian University of Science and Technology about economic valuation of social networks. To make my results better, I need information about some elements describing social networks:

1. Number of members in the social network
2. Average content created per day/month (what kind of content created depends on the nature of the service. e.g. pictures uploaded, blogs created etc.)
3. Average number of connections/friends for a user in your network
4. Estimated value of your network in USD

If you are able to answer some (or all of these questions) for your social network, it would substantiate my results significantly. All answers may be anonymous, if you prefer so.

For any questions, do not hesitate to contact me.

Thank you for your time and help.

Best regards,

Martin Falck-Ytter

B Paper

A paper based on this study, written by Harald Øverby and myself, is attached at the next six pages. The paper has been submitted to the World Telecommunications Congress (WTC) 2012, Miyazaki, Japan [48].

An Empirical Study of Valuation and User Behavior in Social Networking Services

Martin Falck-Ytter and Harald Øverby
Department of Telematics
NTNU
Trondheim, Norway

Abstract— Social networking services (SNS) have emerged as a crucial aspect for private people and businesses worldwide. The usage of SNS in the private sector has seen an exponential use the last years, which has also lead businesses to increase their visibility on SNS. SNS such as Facebook and Twitter have enjoyed huge success due to a large user base and attention from the business sector. The value of a specific SNS is dependent on a number of key characteristics, such as the number of users and the amount of content produced. In this paper we present empirical data for valuation and user behavior in SNS. The data used are from the most common SNS today, including Twitter and Youtube. We provide an overview of how the number of users and content produced influence the total valuation of a SNS. We also develop analytical models capturing the network effects in SNS. The analysis provided can be used to estimate both value and user behavior of future SNS.

Keywords-Social networking services; network valuation

I. INTRODUCTION

The popularity and use of Social Network Services (SNS) has exploded during the last decade. SNS such as Facebook, Twitter and MySpace have millions of followers, many who use a lot of their daily hours updating their profiles and posting new entries. As an example, Facebook has now over 500 million users, in which 50 % log on every day [1]. This evolution has lead to an increased focus from the business sector on SNS, as a mean to reach potential customers.

Several network laws have been proposed to model either user behavior or value. Examples of these laws include Sarnoff's law for broadcast network valuation, Metcalfe's law for valuation of communication networks and Zipf's law for estimating popularity of content. However, the validity of these laws is uncertain for SNS. Behavior laws have to a small extent been verified within the SNS domain. Similarly, most common network valuation laws are based on a theoretical approach. It is therefore unclear how precise they are for SNS valuation.

In this paper we will study empirical data regarding user behavior in SNS. Adjusting the exponent in the Zipf probability mass function, we calculate the best-fit function for popularity of Twitter members and Youtube videos. We see whether content productivity increases with network size and present a response surface model for SNS valuation. The

contributions in this paper are empirical findings clarifying user behavior in SNS further and a SNS valuation formula.

The rest of the paper is organized as follows: Section II presents a background on SNS. Section III presents related works on valuation of networks and SNS in particular. Section IV outlines our modeling and data gathering approach. Section V contains results. Section VI presents key findings in a conclusion.

II. SOCIAL NETWORKING SERVICES

SNS are online platforms for social interactions. With their presence, SNS enabled new services such as sharing of media, event planning and creation of interest groups instantly available for worldwide consumption. The popularity of SNS has not remained unnoticed. A large proportion of total web traffic today is generated by the largest SNS. According to the Internet traffic monitoring company Alexa, around 40 % of global Internet users visit Facebook daily [2]. Some of largest SNS today include LinkedIn (launched 2003), Facebook (launched 2004), Youtube (launched 2005) and Twitter (launched 2006).

In SNS, the variety and quality of content is a key factor for success. Encouragement to produce content is therefore commonly observed in these networks.



Figure 1: Visualized connections on Facebook (image from Facebook)

III. NETWORK VALUATION

Several network laws have been proposed to either model behavior or estimate value. The following section presents and discusses the most common network laws proposed.

Sarnoff's law is attributed to the American radio and television pioneer David Sarnoff. The law states that the value of a broadcast network is proportional to the number of subscribers. The reasoning behind this is that the bigger audience, the more you can charge for advertisements in the network. Examples of broadcast networks where the law is applicable include newspapers, radio and television networks.

$$S(n) = n \quad (1)$$

Sarnoff's law is widely accepted as valuation for broadcast networks, but also limited to this network type.

Metcalf's law states that the value of a network of n compatible communication devices is equal to n^2 . The law can be understood mathematically as the number of possible links in a communication network: each of the nodes in a network of size n can reach $n-1$ nodes. This gives $n(n-1)$ links. But a link from node A to node B , is the same as the link from node B to node A . Therefore; the total number of unique links is equal to:

$$M(n) = \frac{(n-1)n}{2}$$

$$M(n) \approx n^2 \quad (2)$$

Metcalf's law assumes that all network connections are of equal value to an individual user. This can obviously not be true for all network sizes. It is impossible that all users connected to a large network will provide equal value to each other, if any value at all. Aspects like culture, religion and geography affect the utility derived from connections in a network.

In [3], David R. Reed argues that there are some network structures where network value can scale even more than Sarnoff and Metcalf's law. He introduces the concept Group-Forming Network (GFN) as a new network category that enables affiliations among subsets of members. Examples of such networks may be chat rooms and online auctions. Reed defines value as potential connectivity for transactions, which for a GFN is equal to the potential number of subgroups. In a network of n members, each element can be included or not in a subgroup. This gives 2^n possible subgroups in total. However, this includes two non-proper subsets: one where no elements are included and n sets where only one element is included. Therefore, according to Reed's law, the value of a GFN is equal to:

$$R(n) = 2^n - n - 1$$

$$R(n) \approx 2^n \quad (3)$$

Since Reed's law grows even faster than Metcalf's law, it is vulnerable for the same criticism. However, it is important to highlight that Reed talks about value of potential and not actual affiliations. This fact makes the law unpractical for real network valuation.

In [4], the authors accuse Metcalf and Reed for overestimating the value of networks. They argue that the main fundamental fallacy underlying Metcalf and Reed's law

is the assumptions that all potential connections or subgroups are of equal value to a network member. They reason that, since some connections are not used at all and some very rarely, an equal assignment of value to each connection or group is not justifiable. They suggest a new way to value a general communication network of size n . Based on Zipf's law, Tilly and Odlyzko argue that the value of a user scales as $\log(n)$. This leads to a total network value, in a network with n members, of:

$$T - O(n) = n \log(n) \quad (4)$$

Even though Tilly-Odlyzko's law seems to be able to describe real world observations of network effects, there are some downsides with the law. In their reasoning, Tilly and Odlyzko assumed that a network member derives value according to Zipf's law. However, Zipf's law is intended to describe popularity, not value. Whether this approximation is justifiable remains unclear. In addition, Odlyzko and Tilly did only provide some examples where Zipf's law could be an accurate describer of popularity. Whether the law is a good estimation of popularity in all networks remain unanswered. As we later shall see, it is also important to estimate the exponential value in Zipf's law. Without the exponential value specified, the function might differ very much; as the only restriction is that the value is greater than 0.

In [5], Rod Beckstrom proposed a new model for network valuation. According to Beckstrom, the model can be used to value any network type and size. In this model, the present value of any network is equal to the sum of the net present value of the benefit of all transactions minus the net present value of the cost of all transactions. Note that transactions only are carried out if the benefit is higher than the cost of the transaction. All values are discounted over any given period of time. In mathematical notation, Beckstrom's law is formulated as¹:

$$\sum_{i=1}^N V_{i,j} = \sum_{k=1}^M \frac{B_{i,k}}{(1+r_k)^{t_k}} - \sum_{l=1}^P \frac{C_{i,k}}{(1+r_l)^{t_l}} \quad (5)$$

Although Beckstrom's law may give correct results, it introduces a new problem: How are you going to get the beneficial value and cost of every transaction in a network? This question must be as hard to answer as the original problem: How valuable is a network?

Zipf's law is named after George Kingsley Zipf and refers to the fact that several types of data follow a Zipfian distribution. If k is the rank of elements from a data set (where $k = 1$ is the most frequent data), Zipf probability mass function predicts that out of a population of N elements, where s is the value of the exponent, the frequency of elements of rank k is:

$$f(k, s, N) = \frac{1/k^s}{\sum_{n=1}^N 1/n^s}, s > 0 \in \mathbb{R}, n \in \mathbb{I} \quad (6)$$

Zipf's law has proven to be very accurate for modeling popularity of data, such as words in the English language [6] and sizes of large cities [7]. In [8], the authors showed that

¹ The original paper has some typos. In this paper, r was changed to r_k in the formula and 1 to 1 under the explanation of t_k and t_l

income, web page links and traffic to sites follow a power law distribution. A similar study was performed in [9], where the authors showed that a great number of Internet features follow a Zipfian distribution:

- The level of routers transmitting data from one geographic location to another
- The content of the World Wide Web
- How individuals select the websites they visit and form peer-to-peer communities

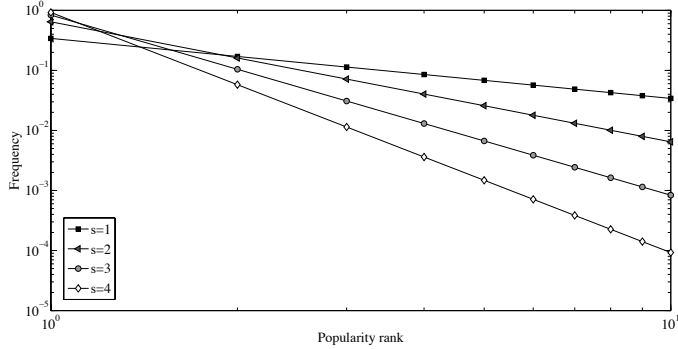


Figure 2: Zipf probability mass function

IV. MODELLING AND DATA GATHERING

A. Popularity of content produced by users in a SNS

Popularity of Twitter users and Youtube videos were compared with Zipf's law to see whether Zipf's law were an accurate describer of popularity in these two social networks. The data gathering approach for the two networks follows:

• *Twitter*

An Internet page containing statistics for the 10 020 most popular Twitter users was used to retrieve data [10]. A Python script was written to gather information about number of followers for each of the 10 020 most popular users².

• *Youtube*

Similarly, a Python script was used to collect statistics from a site containing a list of the 160 most viewed Youtube videos [11].

When these statistics were retrieved, the value of the exponent, s , in Zipf's law was optimized to find the best-fit Zipf probability mass function. This optimization was done with the Levenberg-Marquardt algorithm. The corresponding Coefficient of Determination (R^2) value was calculated with the following formula:

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^n (y_i - f_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (7)$$

Where SS_{err} is the residual sum of squares, SS_{tot} is the total sum of squares, y_i are actual observations, f_i are estimated values by the regression model and \bar{y} is the average value of y_i . The R^2 value compares the variance of the model's predictions with the total variance of the data.

When best-fit functions were given, studentized residuals were calculated, as they have zero mean and unit standard deviation. This makes it possible to determine how far an observation is away from the mean in terms of standard deviation units. Studentized residuals does also compensate for the leverage effect. Therefore, it is easier to observe outliers. The studentized residuals were compared against the following assumptions for non-linear regression:

- **Plausibility:** The regression model is scientifically plausible.
- **Normality:** The variability of values around the curve follows a Gaussian distribution.
- **Homoscedasticity:** The response variables all have the same variance.
- **Accuracy:** The model assumes that you know the independent variable(s) exactly.
- **Independence:** The errors are independent of each other.

B. Content productivity as a function of network size /

C. Valuation of SNS as a function of content and size

To gather information for the two last parts of this study, 203 SNS were visited or requested for the following data:

- Average content produced per day
- Number of members
- Estimated market value in (United States Dollars) USD

The relationship between network size and content created in SNS was studied to see whether content productivity increases with network size. 15 SNS were able to provide the data requested. Best-fit formulas were compared with the data and an F-test was performed on plausible regression models. The F-test compares change in sum of squares with change in degrees of freedom for two models, where one model is more complicated (has more adjustable variables). An F ratio higher than 1 means that either that:

- The more complicated model is correct
- The simpler model is correct, but random errors made the complicated model to fit better than the simpler model.

A P-value describes the probability for the second case to be true. The confidence level for the P-value was in this study set to 0.05, corresponding to a 5% chance of rejecting the correct model. This way it could be concluded whether one model significantly fitted the data better.

Three alternative response surface models for valuation of SNS were calculated and compared. The models were based on two variables: content created and network size. Unfortunately, only 5 social networks were able to provide the information needed. The software Mathematica 8 was used to calculate best fit for a linear, quadratic and power response surface. Any plausible response surfaces were converted, based on the function estimated in the previous section, to a

² 10 011 observations were made, as some entries were zero or out of order

function only dependent on network size. This way, it was possible to compare the response surface model against network effects and network laws presented earlier in this study.

V. RESULTS

The empirical results from this study is presented below:

A. Popularity of content produced by users in a SNS

1. Twitter

The resulting data from the script were retrieved 4.26.2011 and are given in figure 3.

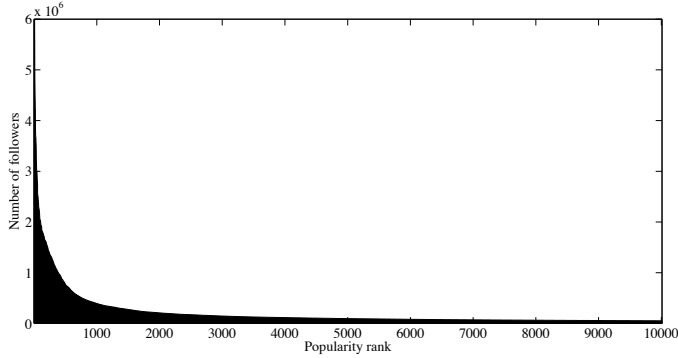


Figure 3: Number of followers for the 10 011 most popular Twitter users

To be able to compare the popularity of Twitter users with Zipf's law, the data were transformed to frequency:

$$f_k = \frac{n_k}{2147851407} \quad (8)$$

Where n_k is the number of followers for Twitter user with popularity rank k and 2147851407 is the total number of followers for the 10 011 most popular Twitter users. The following minimum sum of squares problem was solved to find the optimal value of the exponent, s :

$$\min \sum_{k=1}^{10011} (f_k - \frac{1/k^s}{\sum_{n=1}^{10011} 1/n^s})^2 \text{ subject to } s > 0 \quad (9)$$

The value s for the minimal sum of squares was 0.56. In figure 4, Zipf's law with $s=0.56$ is plotted with the data from Twitter.

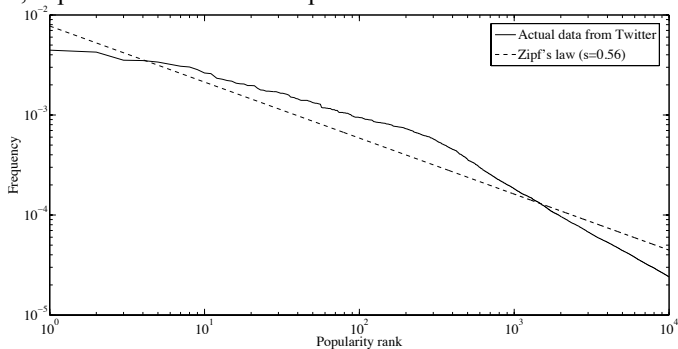


Figure 4: Zipf's law plotted against popularity of Twitter members

Studentized residuals and histograms of studentized residuals are given in figure 5 and 6.

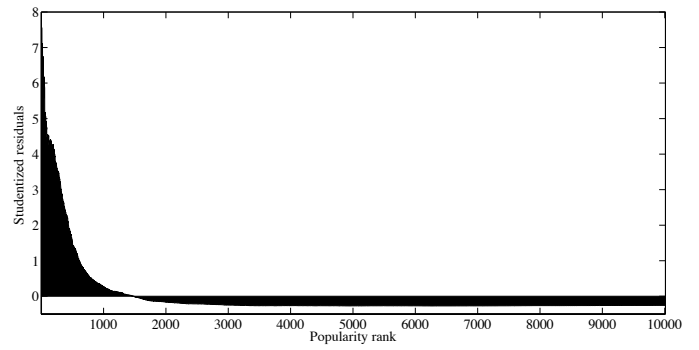


Figure 5: Studentized residuals after the fitting (Twitter)

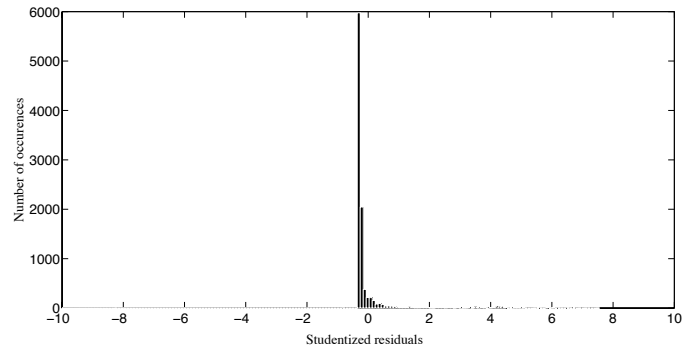


Figure 6: Occurrences of studentized residuals (Twitter)

The result of the fitting where then tested against the regression assumptions listed earlier to test wheter Zipf's law was an accurate describer. The data from Twitter fitted with Zipf's law ($s=0.56$) had an R^2 value of 0.8481. Even though this value does not indicate a very good fit, it does not mean that it is scientifically implausible to fit Zipf's law with the data from Twitter. The occurrences of studentized residuals in figure 6 are not mirrored around origo. The normality assumption is therefore violated. At the most extreme, the standardized residuals were as high as seven times the standard deviation unit (figure 5). This implies that the homoscedasticity property is violated. The independent value, k , is known exactly, so this property is not violated. The last regression assumption is independence. There is a systematic pattern in the studentized residuals in figure 5. This implies that the studentized residuals are not independent of each other. This last property is therefore also violated. As we have seen, several regression assumptions are violated. This implies that the data from Twitter cannot be fitted accurately with Zipf's law, at least not for the whole interval examined.

2. Youtube

Number of views for the 160 most popular Youtube videos was retrieved 4.27.2011 by the script and is given in figure 7. The procedure performed was the same as with the fitting of Zipf's law with Twitter. The optimal value of the exponent, s , was 0.45. The results from the fitting are illustrated in figure 8, 9 and 10.

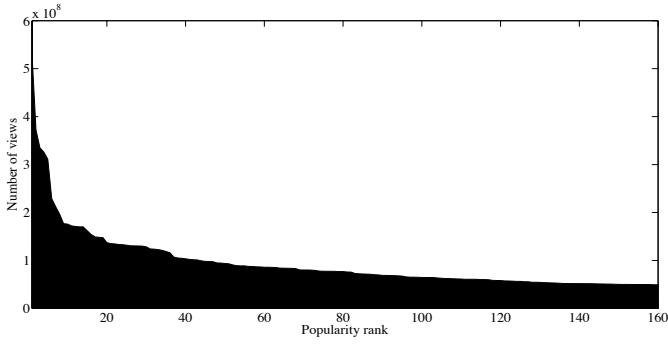


Figure 7: Number of views for the 160 most popular Youtube videos

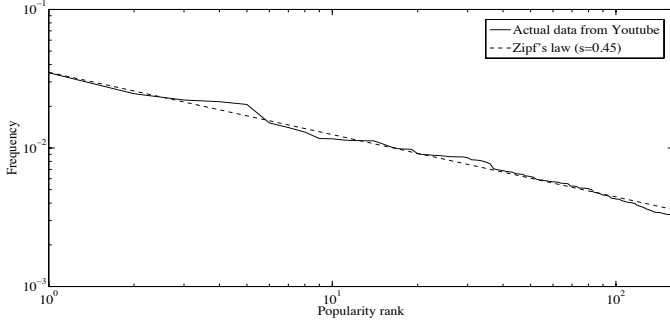


Figure 8: Zipf's law plotted against popularity of Youtube videos

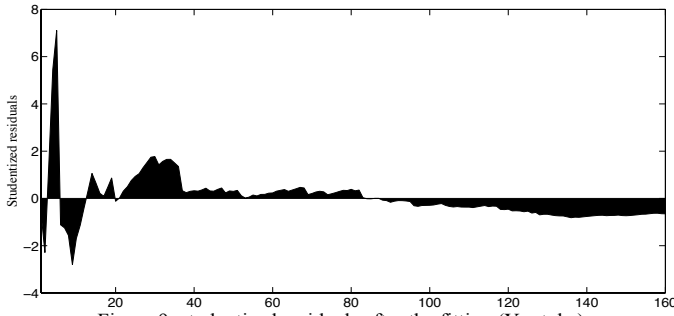


Figure 9: studentized residuals after the fitting (Youtube)

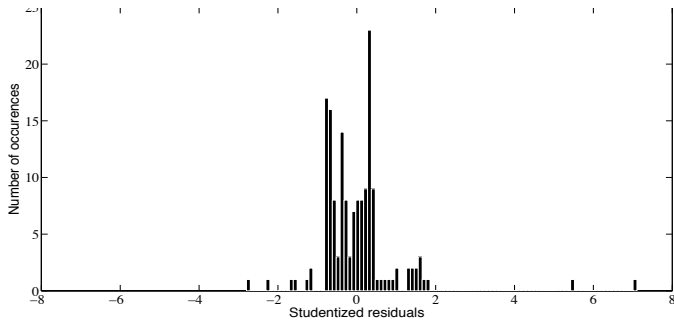


Figure 10: Occurrences of studentized residuals (Youtube)

Tested against the regression assumptions listed earlier, Zipf's law fitted the data from Youtube with a very high R^2 value (0.9859). This means that the regression line almost perfectly fits the data. It is therefore very likely that Zipf's law is a good describer of popularity of YouTube videos. The occurrences of studentized residuals are distributed approximately as a Gaussian distribution (figure 10). The second requirement is

therefore fulfilled. The variances of the studentized residuals (figure 9) are mostly equally divided in the interval, except for some outliers. These were approximately seven times as high as the unit standard deviation. Overall, the homoscedasticity assumption was preserved with some exceptions. The values of popularity rank are known, so this property is not violated. The last regression requirement, independence, is not perfect, as the residuals are not completely randomly distributed (figure 9). However, a systematic misfit is not registered here. With regard to these points, Zipf's law seems to be a good describer of popularity of Youtube videos.

B. Content productivity as a function of network size

Figure 11 displays scatter plot of content created and network size, obtained as described in chapter IV-B. Best-fit formulas and corresponding R^2 values for a quadratic, linear and power fit are plotted in the same figure.

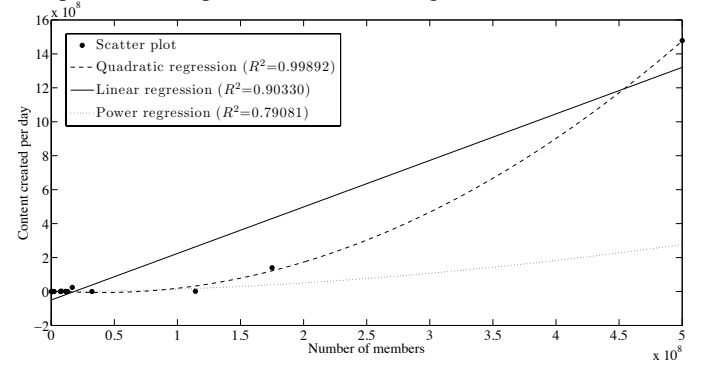


Figure 11: Comparison of best-fit quadratic, linear and power functions

Linear and quadratic functions had high R^2 values and are therefore both plausible estimates. Their best-fit formulas are given below:

$$c(n)_{quadratic} = 7 \times 10^{-9} n^2 - 0.5585n + 5 \times 10^6 \quad (10)$$

$$c(n)_{linear} = 2.7411n - 5 \times 10^7 \quad (11)$$

The P-value for the comparison of a first and second order polynomial was < 0.0001 . Thus, a quadratic fit was statistically significant at confidence level 0.05. This means that the quadratic model fitted the data significantly better than the linear model. Consequently, average productivity increased with network size for SNS studied.

C. Valuation of SNS as a function of content and size

The best-fit linear and quadratic response surfaces had both undesirable properties, as they both had negative value estimations and value decrease after increase in either network size or content productivity. These models were therefore considered inappropriate. The best-fit power response surface was:

$$V_{prs}(n, c) = 14.1514n^{0.892437}c^{0.167022} \quad (12)$$

Where n is the network size and c average content created per day. Details about the best-fit power response surface function are given in table 1.

Table 1: Power response surface regression calculations

Network size	Average content created per day	Actual value	Estimated value	Actual/estimated ratio	Residual
500000000	14785000000	4100000000	40988874992	1.000271	-11125007
175000000	95000000	770000000	6912687207	1.113893	-787312793
10000000	5840	9000000	106397693.5	0.845882	16397693
3000000	2299843	9000000	98574741.69	0.913012	8574741
114270752	142609	30000000	1595346308	0.188046	1295346308
1800000	562500	9500000	49389314.69	1.923492	-45610685
Residual Sum of Squares					2.3×10^{18}

Figure 12 displays the response surface estimated by $V_{prs}(n,c)$.

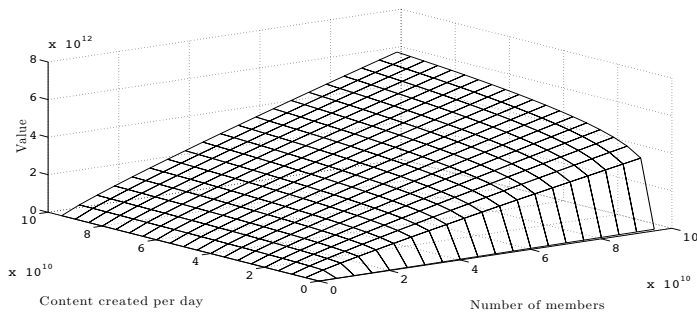


Figure 12: Power response surface for estimated network value

It was earlier in this study concluded that content as a function of network size grew quadratically. If we in asymptotic terms, substitute content productivity in $V_{prs}(n,c)$, we get a model describing network value as a function of network size:

$$V_{prs}(n, c) \approx n^{0.892437} c^{0.167022} \quad (13)$$

$$V_{prs}(n) = 14.1514n^{0.892437}(n^2)^{0.167022} \quad (14)$$

$$V_{prs}(n) = n^{1.226481} \quad (15)$$

$V_{prs}(n)$ compared against network laws proposed is given in figure 13.

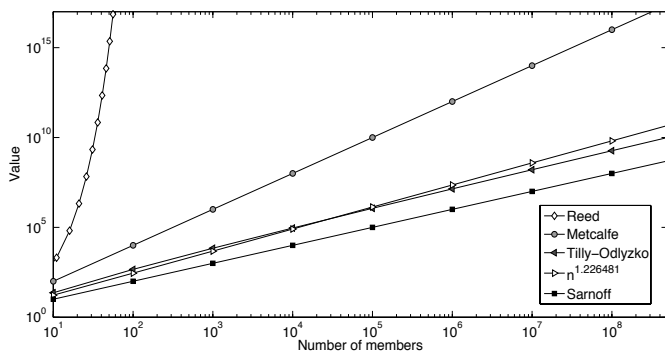


Figure 13: Comparison of results in this study and proposed network laws

VI. CONCLUSION

In this paper we have presented analytical models for user-behavior and SNS valuation with the following key findings:

- Zipf's law was not an accurate describer of popularity of Twitter members.
- Zipf's law was a good describer of popularity of Youtube videos.
- Content productivity increases with network size for SNS studied.
- An empirical model for SNS valuation was proposed based on two variables: network size (n) and average content created per day (c). The best-fit response surface was the following power function:

$$V_{prs}(n, c) = 14.1514n^{0.892437} c^{0.167022}$$

- Compared to proposed network laws, the power response surface grows approximately as Tilly-Odlyzko's law in asymptotic terms.

Further work:

- Test Zipf's law against popularity of content in more SNS
- Gather more data (used during the estimation of a power response surface model) to either:
 - Estimate the variables in a power response surface more accurately
 - Test the proposed valuation model

REFERENCES

- [1] <http://www.facebook.com/press/info.php?statistics>. Accessed 6.5.2011
- [2] <http://www.alexa.com/siteinfo/facebook.com>. Accessed 6.5.2011
- [3] David P. Reed (1999), *That Sneaky Exponential: Beyond Metcalfe's Law to the Power of Community Building*, Available at: <http://www.ate.co.nz/networking/reedslaw.htm>. Accessed 6.3.2011
- [4] Andrew Odlyzko, Benjamin Tilly (2005), *A refutation of Metcalfe's Law and a better estimate for the value of networks and network interconnections*. Available at: <http://www.dtc.umn.edu/odlyzko/doc/met-calfe.ps>. Accessed 6.3.2011
- [5] Rod Beckstrom (2009), *A New Model for Network Valuation*, National Cybersecurity Center. Available at: <http://www.beckstrom.com/The Economics of Networks>. Accessed 1.27.2011
- [6] George Kingsley Zipf (1932), *Selected Studies of the Principle of Relative Frequency in Language*, Cambridge, MA.: Harvard University Press.
- [7] George Kingsley Zipf (1949), *Human Behavior and the Principle of Least Effort*. Cambridge, MA Addison-Wesley
- [8] Clay Shirky, *Power Laws, Weblogs, and Inequality*. Availalbe at: http://www.shirky.com/writings/powerlaw_weblog.html. Accessed 2.13.2011
- [9] Lada A. Adamic and Bernardo A. Huberman (2002), *Zipf's law and the Internet*. Available at: <http://access.cs.sci.ku.ac.th/~usa/418591/2010-1/practice/sukumal/Zipf-internet.pdf>.
- [10] <http://twittercounter.com/pages/100/>. Accessed 2.21.2011
- [11] http://www.youtube.com/charts/videos_views?t=a. Accessed 4.27.2011