SNA Projekt

Analyse von Schweizer Newsportalen sowie Fachhochschulen auf Twitter

Gruppe 3
Christian Schneider, Markus Mächler, Kathrin Koebel

HS 2015, 5iCb

Inhalt

Thema

Wir möchten verschiedene Schweizer Newsportale in verschiedenen Aspekten vergleichen. Dabei wollen wir die Popularität und den Einfluss der einzelnen Portale, die Demographie der Nachfolger sowie die Relevanz/Glaubwürdigkeit der Informationen untersuchen.

Fragestellungen und Erwartungen

- Wie viele Followers hat jedes Newsportal?
    - Welche politische Ausrichtung besitzen diese?
- Wie aktiv sind die Portale?
    - Häufigkeit der Tweets
- Sind die User, welche mehreren Newsportalen folgen, auch untereinander vernetzt?
- Sind User, welche dem selben Newsportal folgen, auch untereinander verknüpft?
- Wie aktiv sind die Followers?
    - Likes oder Retweets von Meldungen → Dies könnte Hinweise darauf geben, wie relevant oder glaubwürdig die Informationen sind oder als wie wichtig diese erachtet werden, besonders wenn es sich bei der Aktivität um einflussreiche Followers handelt.
- Wie ist die Demographie der Followers?
    - Wohnort
    - Sprache
    - Geschlecht
- Wie einflussreich sind die Followers?
    - Anzahl Followers
    - Verhältnis zwischen Followers und Followings
- Sind die Gruppen der Followers der einzelnen Portale disjunkt oder gibt es Überschneidungen?
    - Hat die politische Ausrichtung der Newsportale einen Einfluss auf die gemeinsamen Follower?

Mögliche weitere Fragen

Ein weiteres spannendes Gebiet wäre zu untersuchen, was die einzelnen Portale zu aktuellen Themen wie #Fluechtlinge, #wahlench15, etc. schreiben. Dabei könnten folgende Werte untersucht werden:

- Anzahl Tweets zum Thema
- Anzahl Likes und/oder Retweets
- Sentimentanalyse der Tweets

Fetcher

Als Datenquelle haben wir das Twitter API verwendet. Das API stellt die öffentlichen Daten der Twitter Benutzer über ein REST API zur Verfügung. Wir haben uns für den Fetcher für die Programmiersprache Java entschieden und verwendeten Twitter4J[1] als Wrapper für das Twitter API. Die Daten werden in unser eigenes Model von Twitter-Benutzer und Graph überführt und von dieser Form via Gexf4J[2] in das XML-Format gexf exportiert.
Unser Fetcher verfügt darüber hinaus über drei wichtige Features:

1. **SQLite Datenbank**
   Die eingesammelten Daten werden in einer SQLite Datenbank persistiert, damit Anfragen an das Twitter API reduziert werden können. Zusätzlich ist so bei einem Unterbruch (z.B. Internet-Verbindung) sichergestellt, dass keine Daten verloren gehen. Der Export ins Gephi-Format kann unabhängig vom Einsammeln der Daten gemacht werden.

2. **Command Line Interface**
   Um zu steuern welche Aktionen von der Applikation ausgeführt werden, verfügt diese über ein Command Line Interface. So gibt es z.B. diese Befehle:
   fetch newsportals
   fetch newsportalFollowers
   export [optional filename]
   analyse SpecificOverlaps

3. **Mehrere API Clients**
   Um die Zugriffs-Einschränkung des Twitter API etwas abzufangen haben wir die Möglichkeit entwickelt mehrere Clients sequenziell laufen zu lassen. Damit kann die Geschwindigkeit der Datensammlung ganz einfach vervielfacht werden. Hat ein Client das Zugriffslimit erreicht, wird automatisch mit dem nächsten freien Client weiter gemacht. Haben alle Clients ihr Limit erreicht, schläft der Thread bis ein Client wieder abfragen machen kann.

Um ein effizientes Zusammenarbeiten sicher zu stellen haben wir den Fetcher mit Git versioniert und auf Github veröffentlicht: https://github.com/maechler/sna

[1] http://twitter4j.org
[2] https://github.com/francesco-ficarola/gexf4j

Datenquellen

Wir haben Twitter-Benutzer und deren Beziehungen untereinander als unsere Daten. Dabei sind wir von Newsportalen und ihren Followern ausgegangen. Von den Followern haben wir weiter wieder die Follower geholt um zu sehen, wie diese untereinander vernetzt sind.
Newsportale und menschliche Twitter-Benutzer verfügen über die gleichen Attribute. Die Unterscheidung wird anhand eines Attributes “type” vorgenommen. Die folgenden Merkmale haben wir für alle Benutzer gesammelt:

- ID
- Benutzername
- Name
- Typ (Mensch oder Newsportal)
- Beschreibung
- Ort
- Sprache
- Dabei seit (Erstellungsdatum des Accounts)
- Anzahl Followers
- Anzahl Followings
- Anzahl Tweets vom Benutzer
- Anzahl Tweets, die dem Benutzer gefallen

Queries

Um die Daten zu erhalten haben wir die folgenden Abfragen an das Twitter API[3] gemacht:

Route	**Beschreibung**
GET followers/ids	Gibt eine Liste der IDs der Follower zurück, es werden bis zu 5000 IDs pro Request zurückgegeben. Damit konnten wir alle IDs der Newsportal-Follower holen und die Beziehungen unter den Followern abbilden.
GET users/show	Gibt die persönlichen Informationen zu einem einzelnen Benutzer zurück. Dieses Query haben wir benötigt um die Informationen pro Newsportal zu holen.
GET users/lookup	Gibt eine Liste von Benutzern mit allen ihren Informationen zurück. Diese Abfrage haben wir benötigt um eine zufällige Auswahl der Newsportal-Follower anhand ihrer ID zu holen.

[3] https://dev.twitter.com/rest/public

Einschränkungen

Wir mussten mit diversen Limitationen umgehen. Zum einen gab es Grenzen bei der Datensammlung, zum anderen bei der Analyse der Daten. Beides führte dazu, dass wir starke Einschränkungen bei den Daten machen mussten.

Twitter API

Das Twitter API hat sehr viele Einschränkungen, die das Sammeln der Daten erschweren. Für die meisten Abfragen gilt ein Limit von 15 Requests pro 15 Min., was darüber hinaus geht führt zu einer “Rate limit exceeded”[4] Exception. Zudem ist die Rückgabe der Daten beschränkt indem z.B. nicht alle Follower zurückgegeben werden, sondern nur 200 pro Request.
Eine weitere Einschränkung ist, dass gewisse Benutzer ihre Follower nicht öffentlich zugänglich gemacht haben. Dies führt bei Twitter4J zu einer “Not authorized” Exception. Glücklicherweise stellte dies kein grosses Hindernis dar, da bei unseren Daten nur etwa 1% der Benutzer diese Einschränkung aktiviert hatten.
Einen Account mussten wir ausschliessen beim Laden der Follower, da dieser über 1.6 Millionen Follower hatte. Es handelte sich dabei um den Twitter-Verified Account[5], der dazu verwendet wird Accounts von wichtigen Personen oder Institutionen zu verifizieren.[6]

Analyse (Gephi)

Gephi unterstützt nur bis maximal 50’000 Datensätze. Die Gesamtmenge der Follower aller Newsportale beträgt rund 825’000. Wir haben uns deshalb entschieden zufällig 5% der Follower (ca. 41’000) auszuwählen. Um das umzusetzen haben wir zuerst die IDs aller Follower der Newsportale geholt. Das geht relativ schnell, da pro Request 5’000 IDs zurückgegeben werden. Dann haben wir 5% ausgewählt und für diese die ganzen Benutzer-Informationen geholt (200 pro Request).
Da wir uns für das Netzwerk der Follower der Newsportale interessierten, haben wir bei den Benutzern nur diejenigen Follower gespeichert, die auch mindestens einem ausgewählten Newsportal folgen.

[4] https://dev.twitter.com/rest/public/rate-limiting
[5] https://twitter.com/verified
[6] https://support.twitter.com/groups/.../119135-about-verified-accounts

Analysen Newsportale

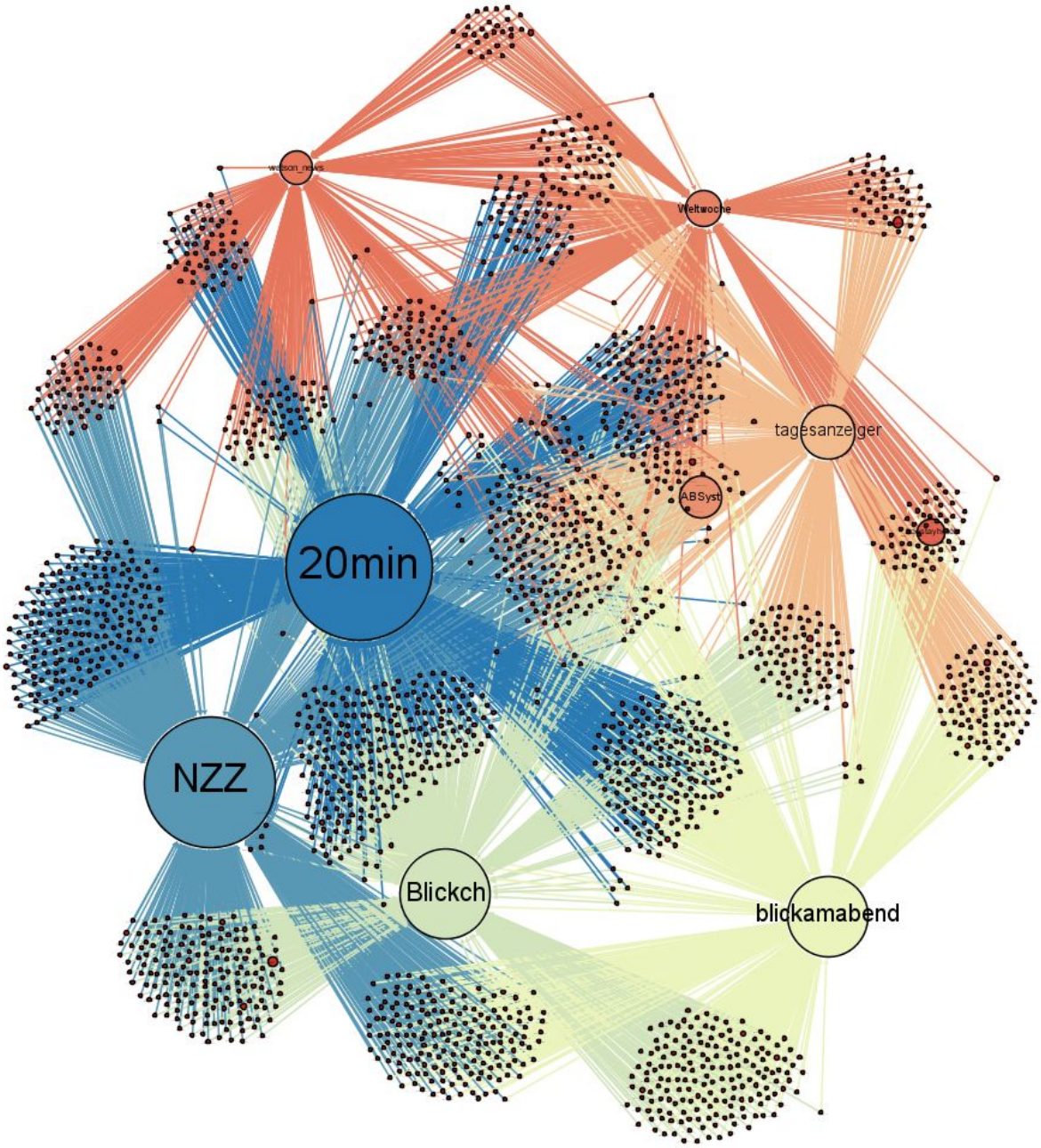
Allgemeine Informationen/Datenanalyse über das Netzwerk

Two Mode Netzwerk (Newsportale und Menschen)
Sieben Newsportale, ca. 37’500 Menschen

Die Reduktion der Datenmenge auf 5% erfolgte nach dem Zufallsprinzip. Dadurch ist es möglich, dass wir besonders aussagekräftige Follower von gewissen Portalen verloren haben und gewisse Metriken verfälscht werden.

Graph

Folgendes Netzwerk besteht aus 2’202 Knoten (Newsportale und ihre Follower) und 4’490 Kanten (Mensch folgt welchem Newsportal). Die Grösse der Knoten stellt den Indegree dar, die Kantenfarbe die Following-Beziehung. Eine Filterung erfolgt zusätzlich indem nur solche Follower dargestellt werden, welche mind. zwei Newsportalen folgen.



Fragestellung:
“Welches Portal hat wie viele Followers?” & “Welche politische Ausrichtung besitzen diese?” & “Wie aktiv sind die Portale (Häufigkeit der Tweets)?”

@tagesanzeiger	76’600 Follower	18’800 Tweets	Linksliberal
@20min	216’000 Follower	42’700 Tweets	Linksliberal
@watson_news	44’200 Follower	23’000 Tweets	Mitte*
@blickamabend	117’000 Follower	28’600 Tweets	Mitte*
@Blickch	130’000 Follower	26’200 Tweets	Rechtsliberal
@NZZ	192’000 Follower	42’600 Tweets	Rechtsliberal
@Weltwoche	49’400 Follower	18’800 Tweets	Rechts

*Annahme

Diese Daten wurden durch manuelles Abfragen bei Twitter (Anzahl Follower) und Recherchen bei Google[7] (politische Ausrichtung) gewonnen.
Es fällt auf, dass die Portale mit den meisten Followern auch am aktivsten sind.

Fragestellung:
“Sind die Gruppen der Followers der einzelnen Portale disjunkt oder gibt es Überschneidungen?”
Die Analyse der gemeinsamen Follower (keine Filterung) ergibt folgende Matrix:

	tagesanzeiger	watson_news	Weltwoche	blickamabend	Blickch	NZZ
tagesanzeiger						
watson_news	43.6%					
Weltwoche	36.5%	36.9%				
blickamabend	44.9%	31.4%	33.5%			
Blickch	45.0%	29.4%	31.7%	64.7%		
NZZ	43.7%	25.3%	27.3%	44.0%	48.5%	
20min	35.0%	21.8%	22.8%	50.2%	57.7%	46.4%

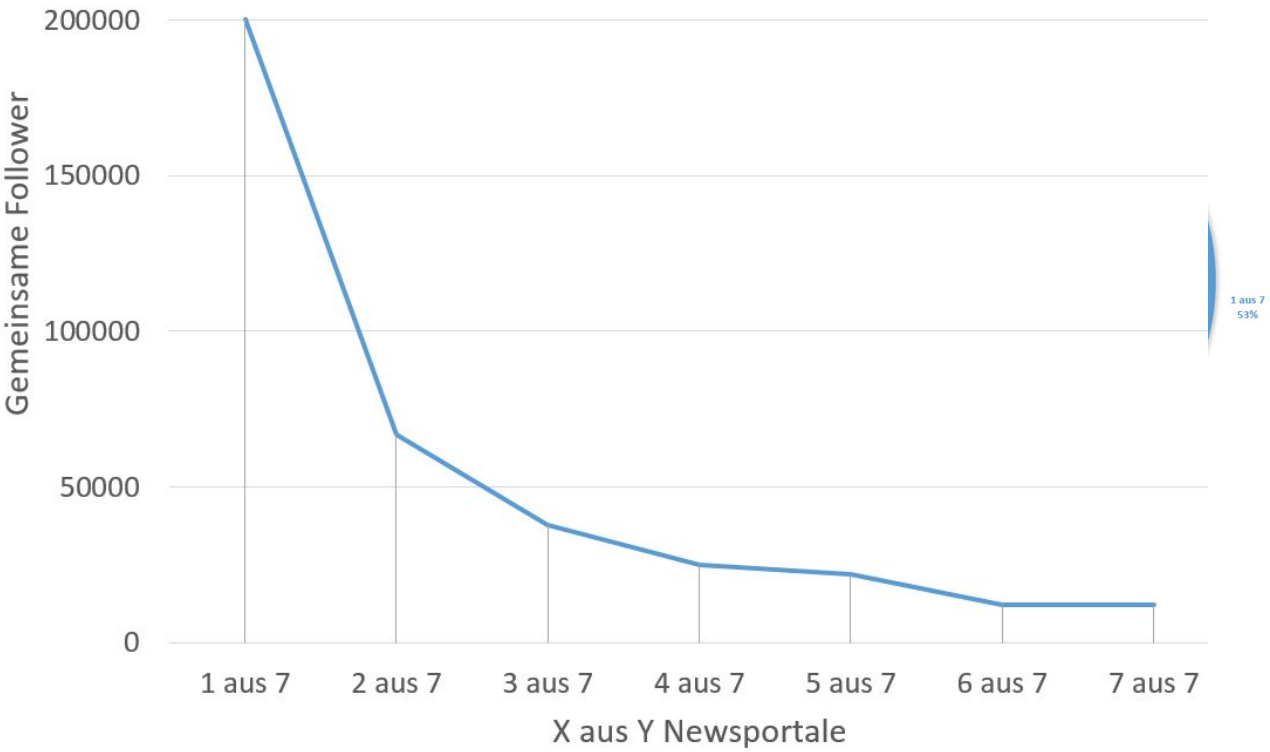
Was gut zu sehen ist, ist dass die Boulevardzeitungen untereinander gut verbunden sind. Gemäss Wikipedia sind die Boulevardzeitungen 20min, Blickch und blickamabend.
Interessant ist die durchgehend hohe Zahl beim tagesanzeiger im Vergleich zu allen anderen Newsportalen.

[7] Quelle: https://de.wikipedia.org/wiki/Medien_in_der_Schweiz

Ein Vergleich über alle 815’591 Follower hat folgende Tabelle ergeben:

Anzahl Follower welche nur 1 aus 7 Portalen folgen:	200’292
2 aus 7:	66’940
3 aus 7:	37’844
4 aus 7:	25’089
5 aus 7:	21’783
6 aus 7:	12’324
Anzahl Follower welche allen von uns ausgewählten Portalen folgen:	12’096

Darstellung der Anzahl Newsportale, denen ein Benutzer folgt als Kuchen- und Liniendiagramm:

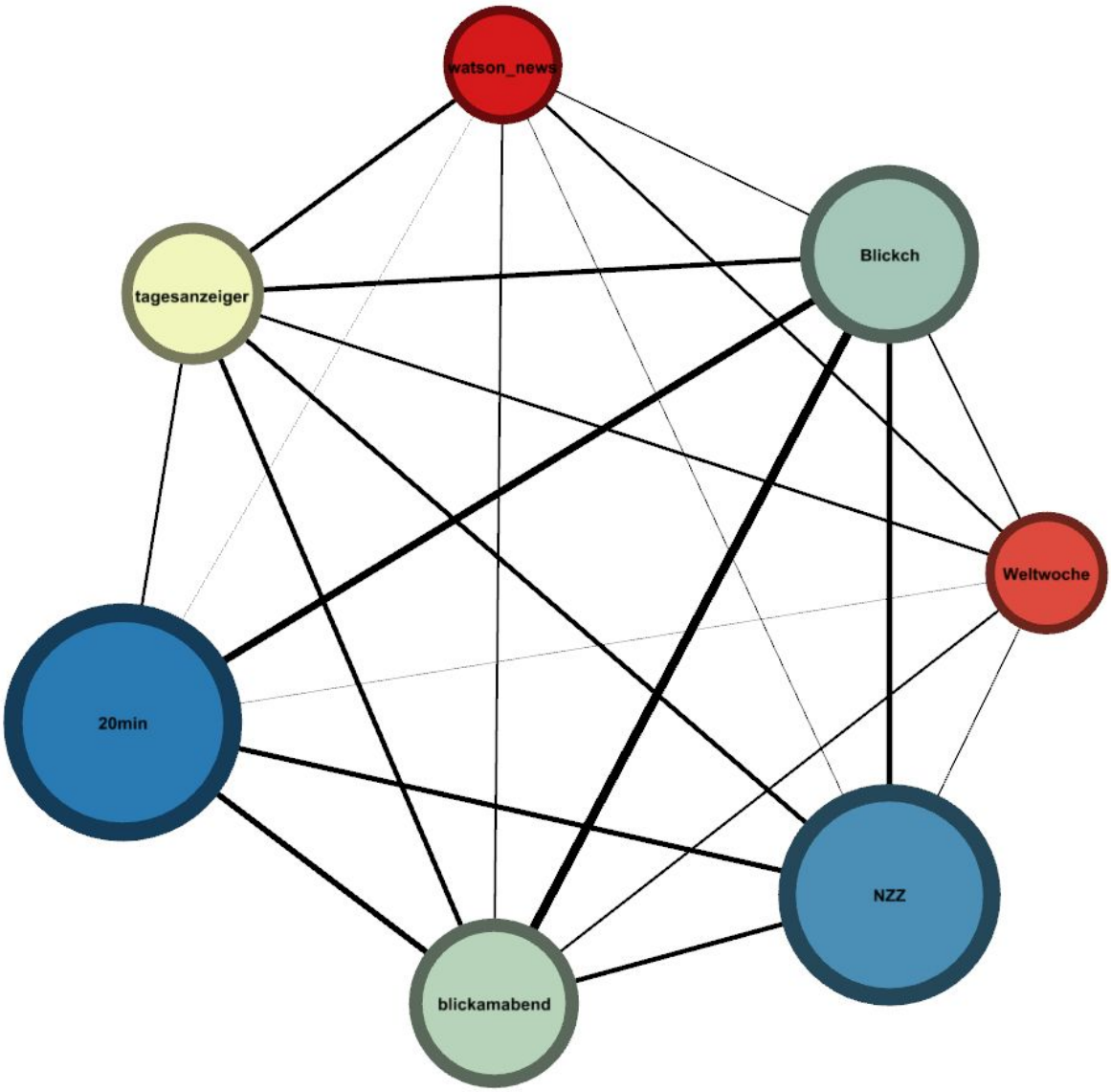


Fragestellung:

“Hat die politische Ausrichtung der Newsportale einen Einfluss auf die gemeinsamen Follower?”

Für die Analyse haben wir einen Graph mit den nicht gefilterten Daten erstellt bei dem die Newsportale gemäss politischer Ausrichtung positioniert wurden.
Knoten: Newsportale. Kanten: Gemeinsame Follower.
Das Kantengewicht wird höher umso mehr gemeinsame Follower es gibt. Das Kantengewicht ist über die Dicke der Verbindung visualisiert.

Interessanterweise sind die gemeinsamen Follower nicht speziell auf eine politische Ausrichtung fixiert, sondern eher auf die Boulevardzeitungen beschränkt. Auch hier schön zu sehen ist die ziemlich gleichmässige Verteilung der gemeinsamen Follower vom tagesanzeiger.

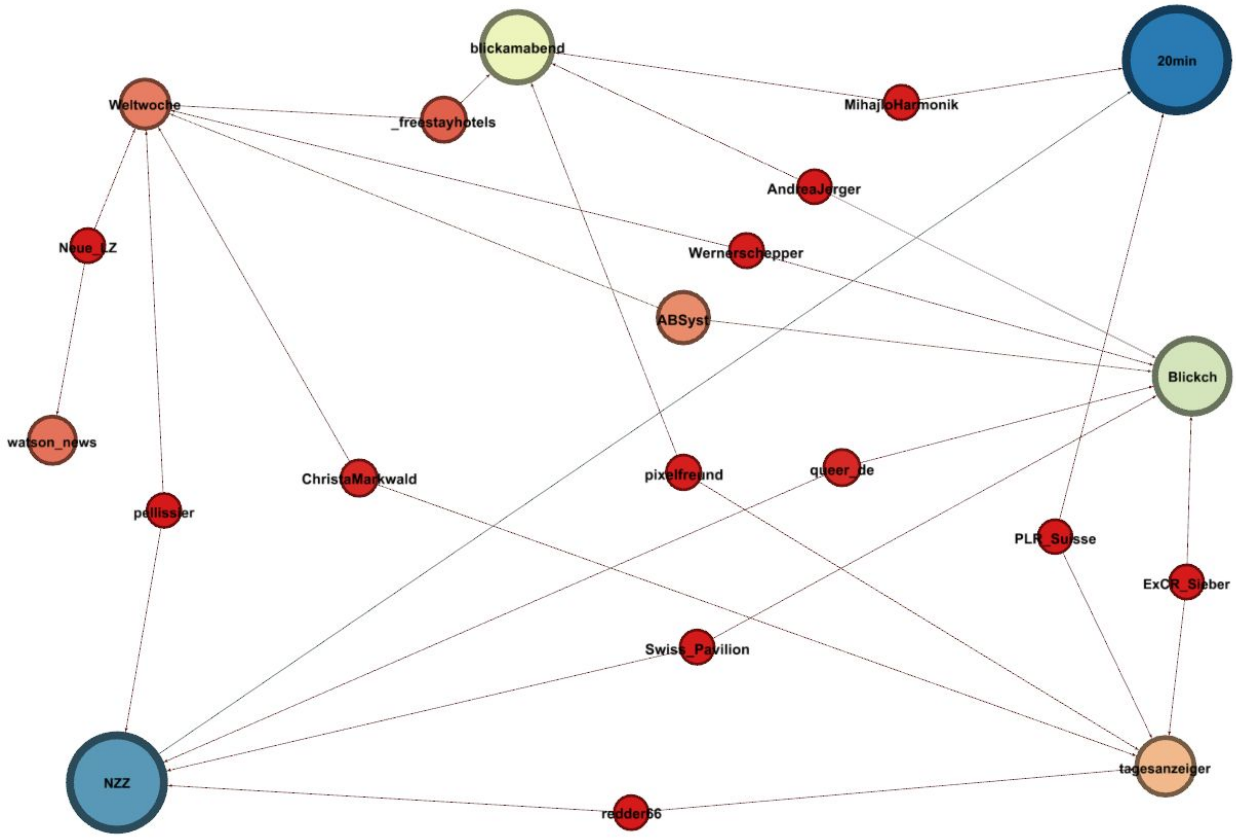


Fragestellung:

“Wie einflussreich sind die Follower (Anzahl Follower, Verhältnis zwischen Followers und Followings)?”

Ein Verhältnis zwischen Followers und Followings ist nicht sehr aussagekräftig. Ein User mit einem Follower und einem Following hat einen Ratio von 1, während z.B. ein User welcher fast niemandem folgt, jedoch von 1000 User gefolgt wird, ein Verhältnis von 0.001 hat.

Folgender Graph stellt alle Twitter-Benutzer dar, welche mindestens 2 Newsportalen folgen und von über 2000 User gefolgt werden. Wir haben hierzu die Auswahl mit den 5% zufällig gewählten Followern genommen.



Fazit: Nur 14 User (bei 5%, hochgerechnet auf 100%: 280 User) und somit 3‰ haben mehr als 2000 Follower)

Folgende Fragen konnten wir nicht beantworten

Für folgende Fragen wäre es nötig alle Follower der Follower der Newsportale zu laden. Diese Datenmenge wäre in der Zeit nicht zu schaffen, weil die Abfragen an das API von Twitter so stark limitiert sind.

- *Sind die User, welche mehreren Newsportalen folgen, auch untereinander vernetzt? (n-Clique? k-Core?)*
- *Sind User, welche dem selben Newsportal folgen, auch untereinander verknüpft?*
- *Wie aktiv sind die Followers (Likes und Retweets von Meldungen)*

Auch die Frage zur Demographie der Follower kann nicht beantwortet werden, da das Geschlecht der User nicht über das Twitter API extrahiert werden kann. Dieses aus der Beschreibung zu extrahieren ist sehr schwer und indes ist es fraglich ob dies viel bringt, da nur rund 22.6% dieses Feld überhaupt gefüllt haben.

Aufgrund der Tatsache dass wir durch zu starke Filterung keine sinnvollen Analysen mehr durchführen können haben wir uns entschieden das Thema zu wechseln.

Analysen Fachhochschulen

Untersuchung von Informatik, Engineering und Multimedia Departementen von deutschschweizer Fachhochschulen

Da die Analyse der Newsportale sich aufgrund der grossen Datenmenge als sehr schwierig herausgestellt hat und durch die Reduktion der Daten kaum mehr Verbindungen zwischen den einzelnen Followern hergestellt werden konnten, haben wir ein weiteres Netzwerk gesucht. Die Beschaffung der Daten sollte sehr ähnlich sein, wie bei den Newsportalen, damit wir unsere bisherige Implementierung verwenden konnten.
Wir haben uns für die Analyse verschiedener Informatik-, Engineering- und Multimedia-Departemente von deutschsprachigen Fachhochschulen entschieden, da es sich zum einen um ein kleineres Netzwerk handelt als dasjenige der Newsportale und damit keine Reduktion der Daten nötig ist. Und wir sind zudem davon ausgegangen, dass hier eine stärkere Vernetzung zwischen den Followern der einzelnen Fachhochschulen besteht, da diese Vernetzung auch in der realen Welt besteht (z.B. Dozenten, welche an mehreren FHs unterrichten). Leider wurden wir erneut enttäuscht. Das generierte Netzwerk der folgenden vier Fachhochschulen zeigte vier praktisch unabhängige Netzwerke auf.

Untersuchte Accounts

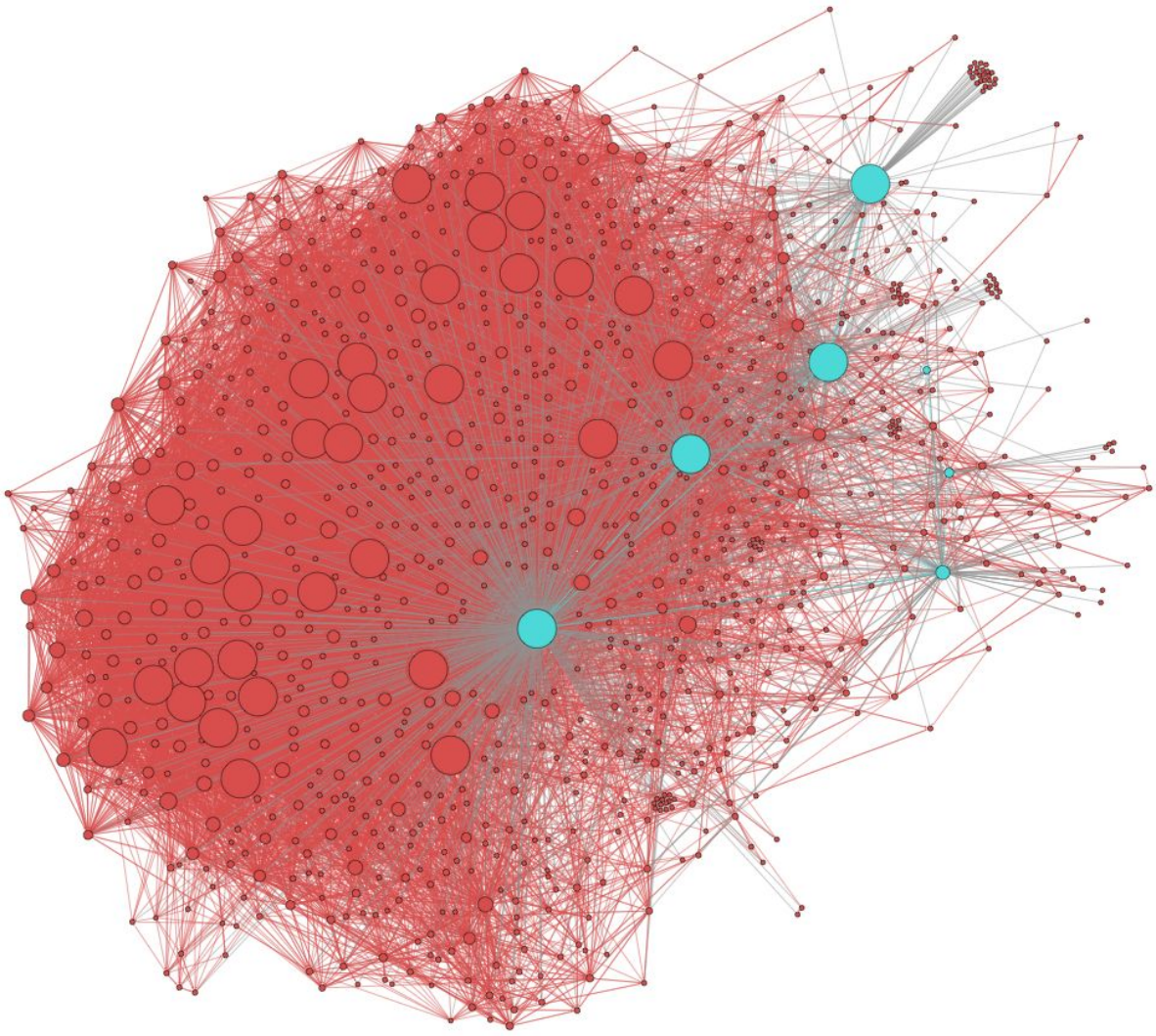
@FHNWTechnik	124 Follower	Hochschule für Technik der FHNW
@engineeringzhaw	459 Follower	ZHAW School of Engineering
@digideation	17 Follower	Studiengang Digital Ideation (Informatik, Design & Kunst), HSLU
@HSR_Informatik	67 Follower	Hochschule Rapperswil Studiengang Informatik

Untersuchung von Twitteraccounts der FHNW aus dem Bereich der Informatik

Aus den oben erwähnten Gründen haben wir den Scope der Untersuchung ein drittes Mal verändert. Diesmal haben wir den Kreis unserer Untersuchung noch kleiner gezogen und uns auf Twitteraccounts der FHNW aus dem Bereich der Informatik fokussiert. Mit diesem Auswahlkriterium sind wir endlich auf ein Netzwerk gestossen, in dem eine stärkere Vernetzung und damit eine spannendere Grundlage für die Analyse bestand.

Untersuchte Accounts

@FHNWTechnik	127 Followers	Hochschule für Technik FHNW
@fhnw_i4ds	65 Followers	Institut 4D Technologien der FHNW
@IT_FHNW	237 Followers	Corporate IT @ FHNW
@iwifhnw	710 Followers	Institut für Wirtschaftsinformatik (IWI) der Hochschule für Wirtschaft FHNW
@ic_fhnw	19 Followers	Studiengang iCompetence FHNW
@ITHGKFHNW	120 Followers	IT Team der Hochschule für Gestaltung und Kunst der FHNW
@dotFHNW	34 Followers	FHNW .Net courses



Die Abbildung zeigt das entstandene Netzwerk, das eine hohe Vernetzung aufzeigt. Die blau hervorgehoben Knoten sind die sieben untersuchten FHNW-Accounts. Bereits in dieser initialen Visualisierung vom Netzwerk zeichnet sich ab, dass der Knoten in der Mitte vom @iwifhnw von zentraler Bedeutung ist.

Allgemeine Informationen zum Netzwerk

Dimensionen des Netzwerkes

- 1’035 Knoten
- 17’458 Kanten (gerichtet)
- Two Mode Netzwerk (Portale und Followers)

Dabei wurden von allen 1’035 Knoten folgende Attribute extrahiert:

- Anzahl Followers
- Anzahl Followings
- Anzahl Tweets
- Anzahl gelikte Tweets
- Twittername
- Name
- Beschreibung (Bio)
- Ort
- Sprache
- Erstellungsdatum des Accounts

Bei den 17’458 Kanten handelt es sich um gerichtete Verbindungen, welche aufzeigen, welcher Benutzer welchem der sieben ausgewählten Accounts folgt sowie auch die Beziehungen unter diesen Benutzern aufzeigt.

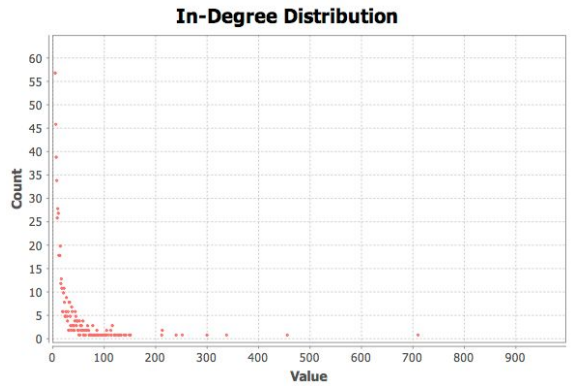
Filterung des Netzwerkes

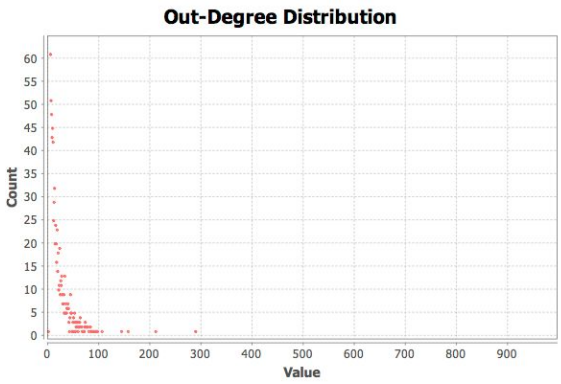
Es wurden 100% der Follower der 7 ausgewählten FHNW Accounts unseres Netzwerkes extrahiert und ausgewertet. Von diesen Follower hingegen wurden jedoch nur die Beziehungen zu anderen Benutzern, welche ebenfalls einem der ausgewählten 7 Accounts folgen, untersucht. Alle anderen Beziehungen der Follower wurden in dieser Analyse vernachlässigt und auch gar nicht im Graph festgehalten.

Graph-Metriken des gesamten Netzwerkes

Degree

Durchschnittlicher Degree (Average Degree) des gesamten Netzwerkes	16.868





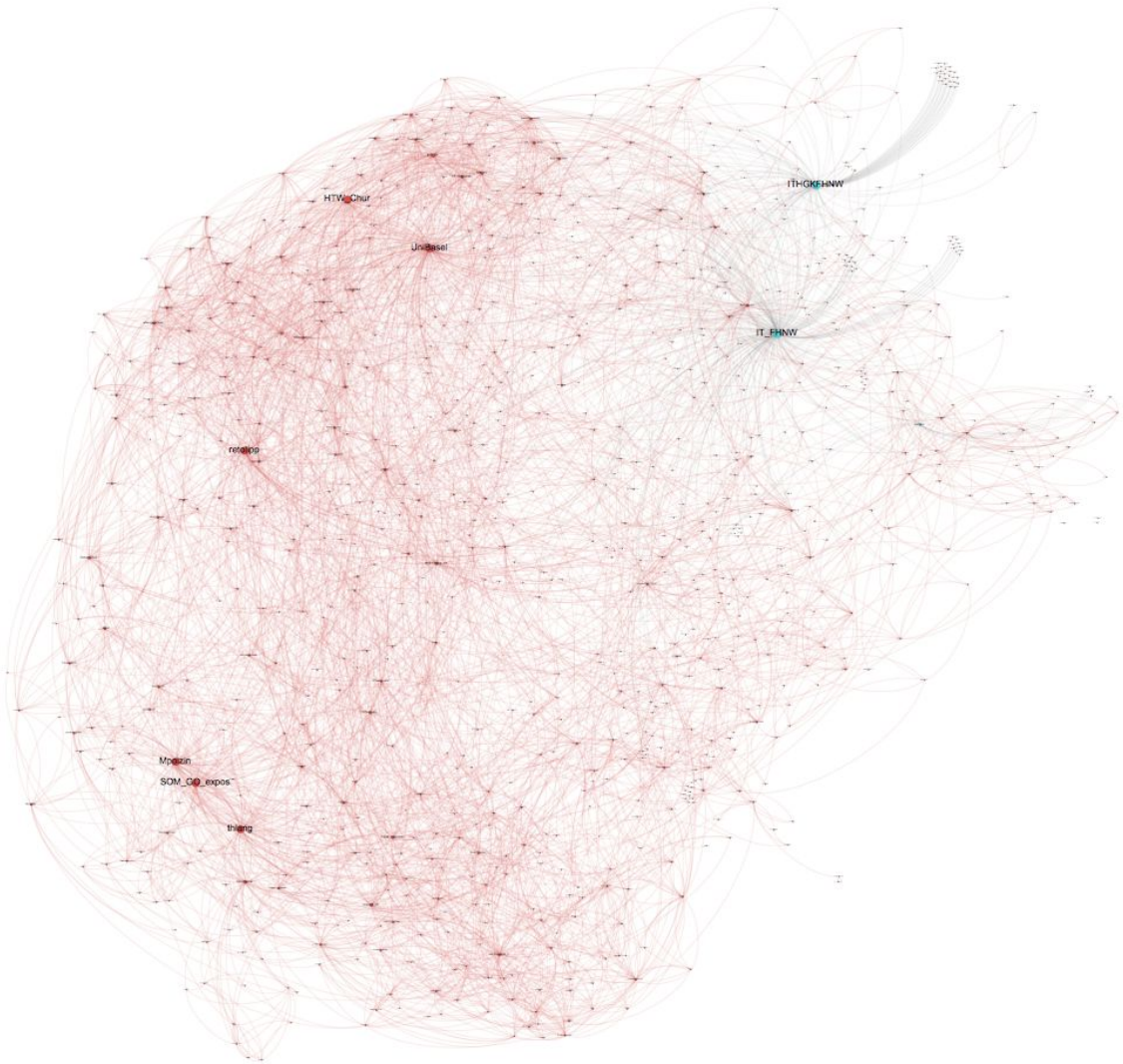
Es scheint, dass der Indegree der meisten Accounts leicht höher ist als deren Outdegree. Das heisst in der Regel haben die Benutzer in unserem Netzwerk mehr Follower als Personen, welchen sie selber nachfolgen.

Betweenness Centrality & Closeness Centrality

Die Betweenness Centrality beschreibt, wie oft ein Knoten auf dem kürzesten Pfad innerhalb des Netzwerkes auftritt.

Durchmesser (Diameter)	9
Durchschnittliche Pfadlänge des kürzesten Pfades zwischen 2 beliebigen Knoten (Average path length)	2.794722165864183
Anzahl der kürzesten Pfade (Number of shortest paths)	808324

Die Closeness Centrality beschreibt wie zentral ein Knoten im Netzwerk positioniert ist. Dies ergibt sich durch den durchschnittlichen Abstand eines Knoten zu allen andern Knoten im Netzwerk.



Die Abbildung visualisiert die Closeness Centrality in unserem Netzwerk. Knoten, welche Teilnetzwerke zusammenhalten, haben oft einen hohen Closeness Centrality Wert, was am Beispiel vom Knoten des ITHGKFHNW in der rechten oberen Ecke schön sichtbar ist.

Eigenvector Centrality

Eigenvector Centrality über eine Iteration	1010.3418079096031

Clustering und Compontents

Durchschnittlicher Clustering Coefficient (Average Clustering Coefficient)	0.380
Graph Density	0.016
Number of Weakly Connected Components	1
Number of Stronlgy Connected Components	252

Da es einen Benutzer gibt, der allen 7 untersuchten Accounts folgt, sind sämtliche Knoten in einem "weakly connected component" (Verbindung sämtlicher Knoten des Graphen ohne Berücksichtigung der Richtung der Kanten) verbunden.

Analyse der Egonetzwerke der 7 ausgewählten Accounts

Degree Centrality/Prestige Indegree

Fragestellung:
"Welcher Account hat wie viele Followers? Wie bedeutend sind die jeweiligen Follower?"

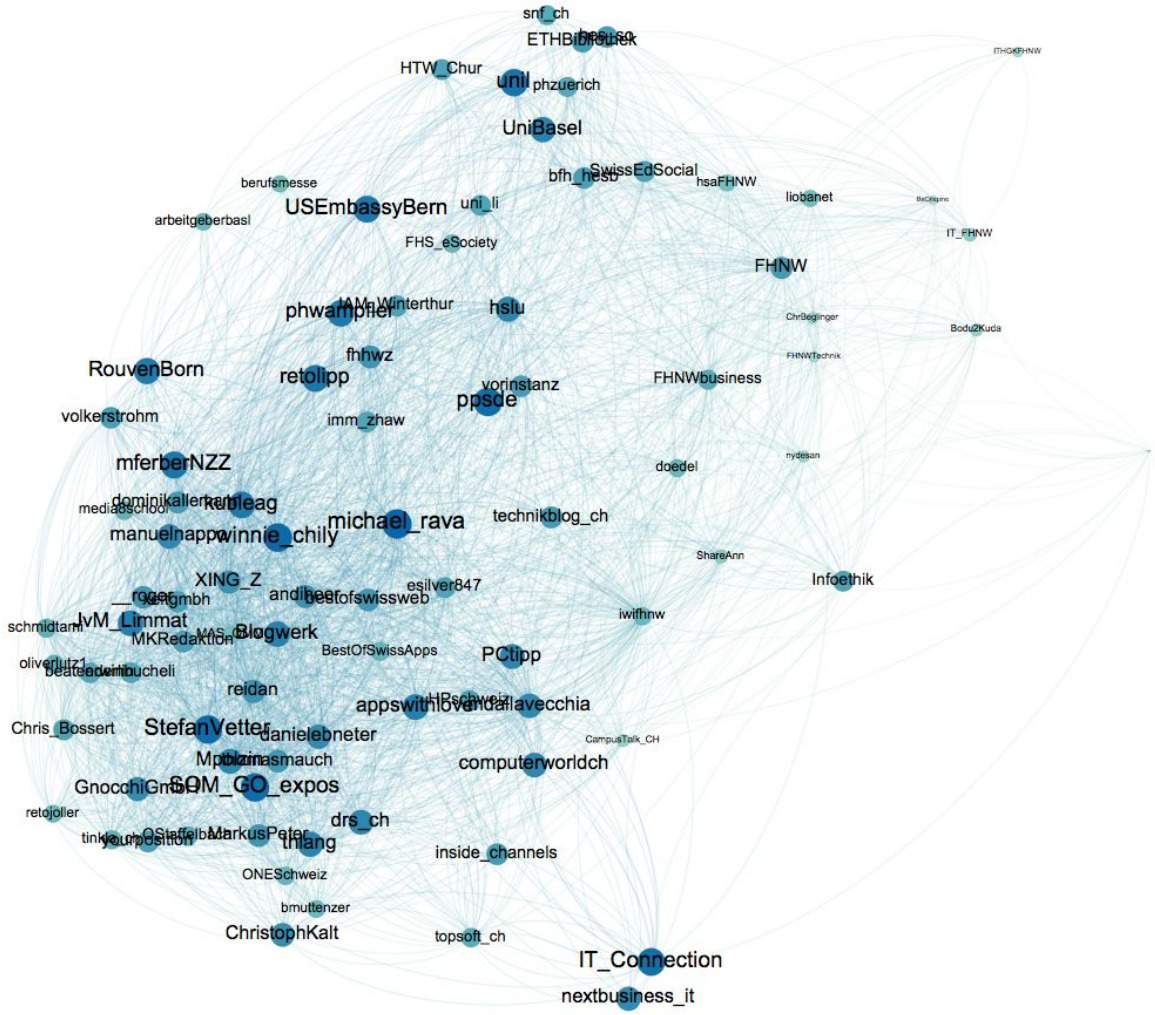
In der folgenden Tabelle werden die Graph Metriken der einzelnen Accounts gegenübergestellt, um ihre Popularität zu vergleichen. Mit der Degree Centrality bzw. Prestige Indegree Metrik kann die Popularität der einzelnen Portale verglichen werden.

Twitter-Account	Anzahl Followers (Prestige Indegree bzw. Degree Centrality)	Average Degree	Anzahl Followers mit mehr als 500 Followers absolut[8]	Anzahl Followers mit mehr als 500 Followers %	Anzahl Followers mit mehr als 1000 Followers absolut	Anzahl Followers mit mehr als 1000 Followers %	Followers /Following Ratio
@FHNWTechnik	135	9.489	52	38.5%	33	24.4%	0.9375
@fhnw_i4ds	65	7.795	29	44.6%	18	27.7%	0.3988
@IT_FHNW	237	8.340	34	14.3%	20	8.4%	4.6471
@iwifhnw	710	18.961	227	32%	146	20.6%	0.9233
@ic_fhnw	19	3.550	4	21.1%	2	10.5%	9.5
@ITHGKFHNW	120	5.372	12	10%	6	5%	120
@dotFHNW	34	3.029	7	20.6%	3	8.8%	2.4286

[8] Bezieht sich auf das Egonetzwerk des jeweiligen Portals, teilweise gibt es Differenzen zwischen Anz. Followers und Nodes im jeweiligen Egonetzwerk

Aus den Werten ist ersichtlich, dass auch einige der kleineren Accounts etliche einflussreiche Followers haben.

Das Followers/Following Verhältnis sagt nicht viel über die Popularität eines Portals aus, da es in der Twitter Community zum guten Ton gehört, dass man sich gegenseitig folgt, da es sich ja um ein “soziales” Netzwerk handelt.[9] Das obenstehende Beispiel zeigt, dass das Follower/Followings Verhältnis demzufolge kein besonders aussagekräftiger Wert ist.

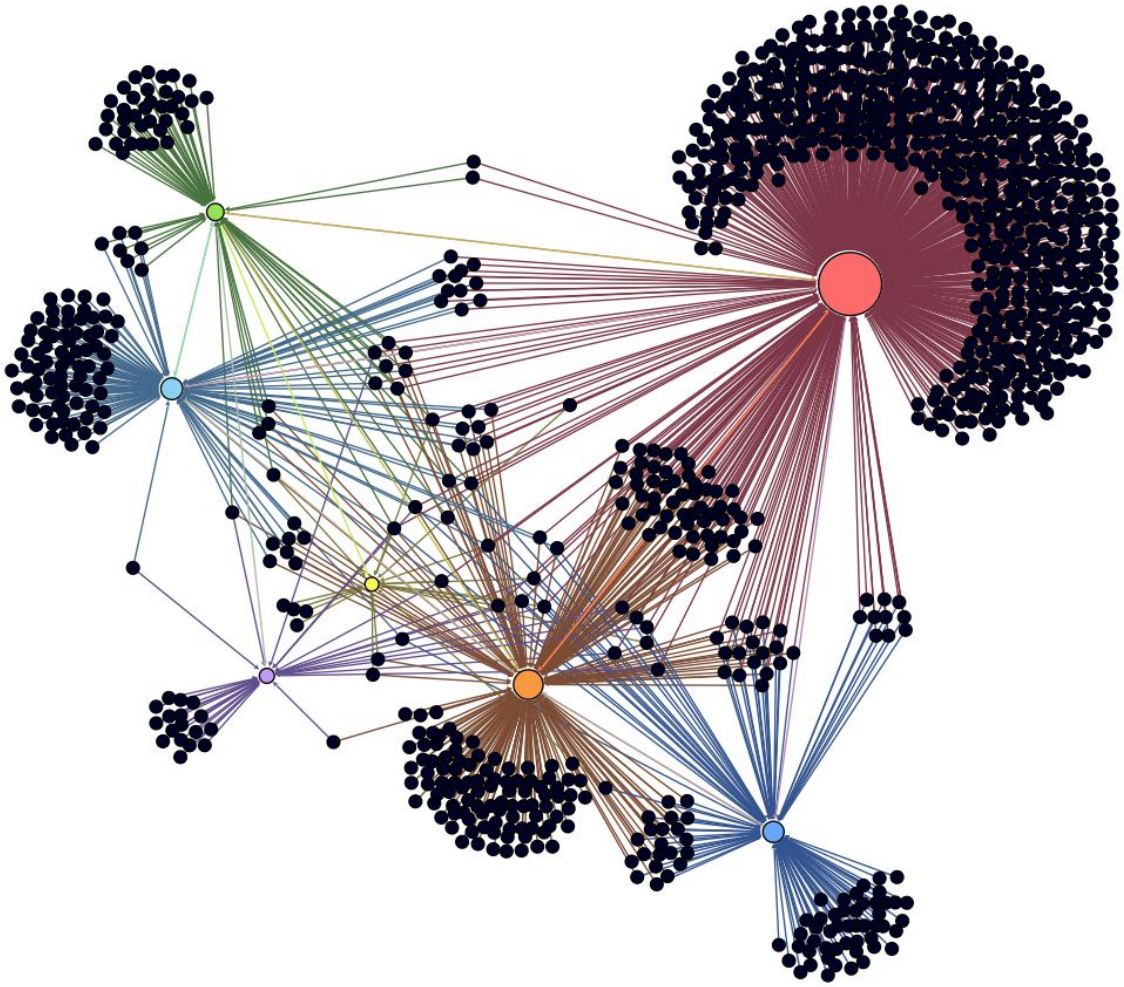


Die Abbildung zeigt die Knoten des gesamten Netzwerkes mit einem Indegree von über 50. Dazu wurden die Anzahl Followers durch die Grösse der Knoten visualisiert.

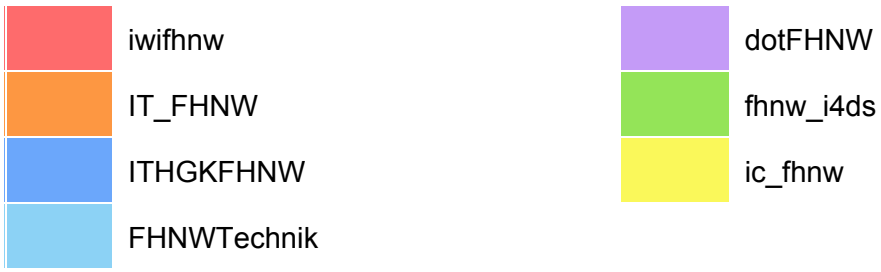
[9] https://deutschtweetor.wordpress.com/2012/05/19/folgen-und-zuruckfolgen-effektiver-followeraufbau

Überschneidungen der ausgewählten Accounts

Fragestellung:
“Sind die Gruppen der Followers der einzelnen Accounts disjunkt oder gibt es Überschneidungen? Sind die User, welche mehreren der von uns selektieren Accounts folgen, auch untereinander vernetzt?”

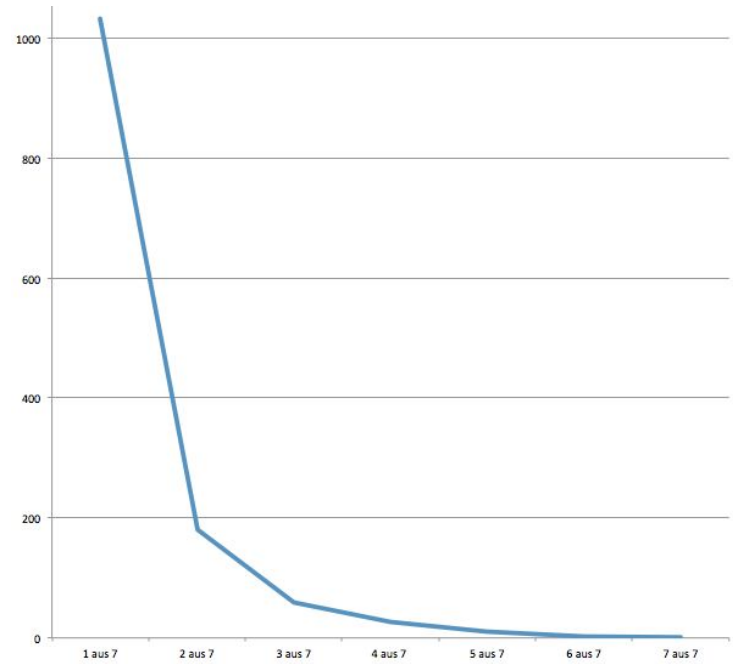


Eine Übersicht aller FHNW-Accounts und ihren direkten Followers. Die Kantenfarbe visualisiert die Following-Beziehung. Die Grösse der Knoten ist relativ zum Indegree.



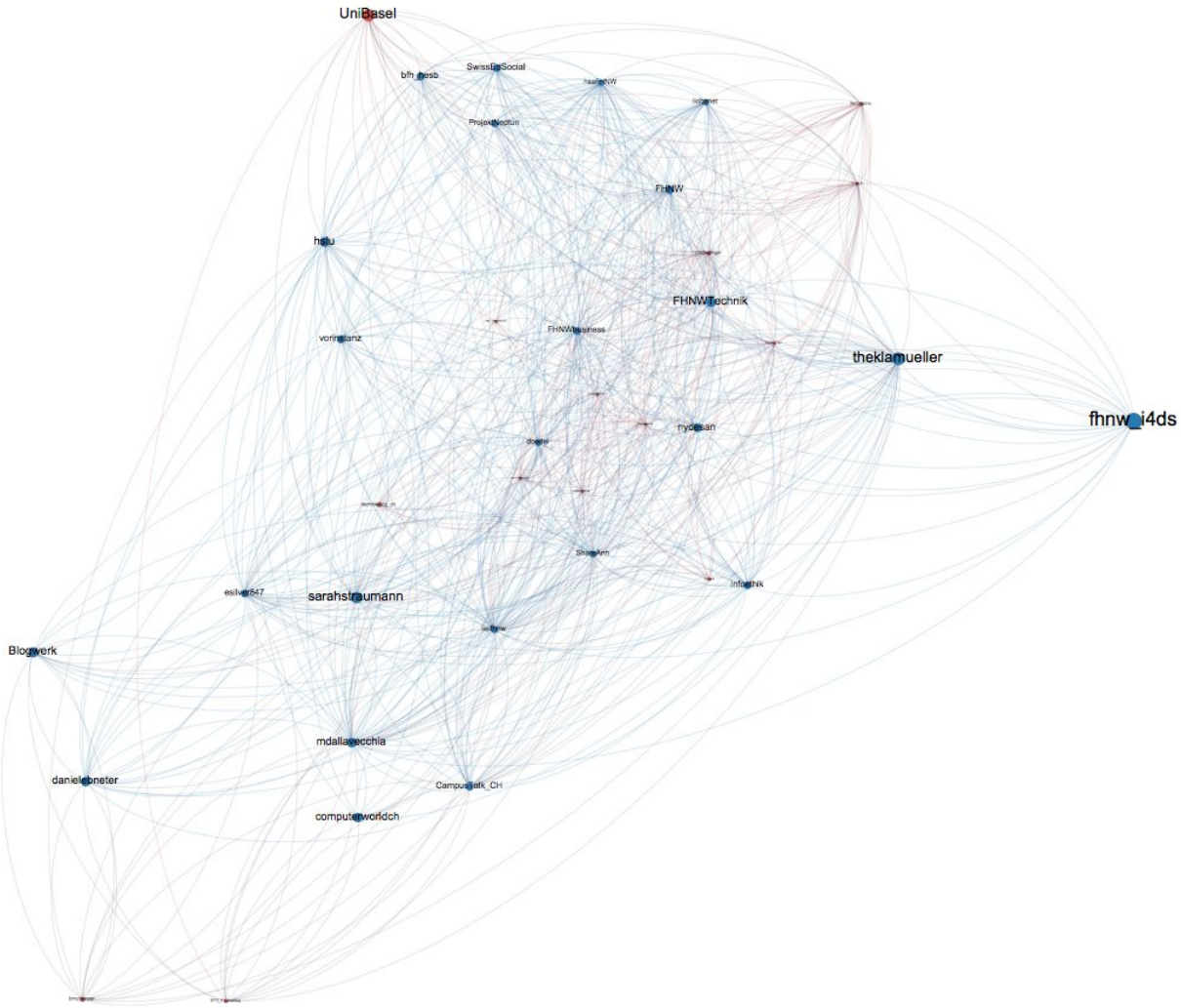
Die Analyse der gemeinsamen Follower der 1’033 eindeutigen (distinct) User unseres Netzwerkes hat folgende Tabelle ergeben:

Anzahl User welche X aus Y der von uns ausgewählten FHNW-Accounts folgen	
1 aus 7:	852
2 aus 7:	123
3 aus 7:	32
4 aus 7:	16
5 aus 7:	7
6 aus 7:	2
7 aus 7:	1



Wer sind die Personen, welche den meisten unserer selektierten Accounts folgen?

7 aus 7	@theklamueller
6 aus 7	@ProjektNeptun, @FHNW
5 aus 7	@fhnw_io, @budimanjob, @fhnw_i4ds, @Bodu2Kuda, @CompasInfoclio, @SwissHigherEd, @School_Expo
4 aus 7	@CHwebdude, @swisseagle71, @schultze_info, @run_sascha, @iwifhnw, @extendance, @FHNWbusiness, @mdallavecchia, @Frau_MINT, @hsaFHNW, @nydesan, @manfredvogel, @FocusIndia2016, @SwissEdSocial, @feuerkaempfer, @ShareAnn



Egonetzwerk von @theklamueller, dem einzigen User, der allen von uns ausgewählten Accounts folgt. Das Netzwerk wurde reduziert auf Knoten mit einer Mutual Degree Range von über 20. Einige der anderen Followers aus der obenstehenden Liste kommen darin vor was darauf hindeutet, dass User, welche mehreren der von uns selektierten Accounts folgen auch untereinander stärker vernetzt sind.

Die folgende Tabelle zeigt die Überschneidungen der Followers von jeweils zwei einzelnen Accounts auf. Die Werte der Schnittmenge sind normiert, wobei 1 die kumulierte Menge aller Followers der beiden Accounts ist. Es ist ersichtlich, dass es bei allen Accounts Überschneidungen der Followers gibt, kein Account hat eine disjunkte Menge von Followern. Die stärkste Überschneidung besteht zwischen @ITHGKFHNW und @IT_FHNW. Dies ist gut nachvollziehbar, da es sich bei beiden Accounts um zwei IT Abteilungen innerhalb der FHNW handelt. Die zweithöchste Überschneidung besteht zwischen @FHNWTechnik und @fhnw_i4ds, was aufgrund der realen Organisationsstruktur ebenfalls naheliegend ist.

@FHNWTechnik		0.2383	0.1913	0.0766	0.0680	0.0741	0.0806
@fhnw_i4ds	0.2383		0.1188	0.0388	0.0714	0.0808	0.0649
@IT_FHNW	0.1913	0.1188		0.2199	0.0934	0.0809	0.2961
@iwifhnw	0.0766	0.0388	0.2199		0.0303	0.0377	0.1111
@ic_fhnw	0.0680	0.0714	0.0934	0.0303		0.2642	0.1295
@dotFHNW	0.0741	0.0808	0.0809	0.0377	0.2642		0.1558
@ITHGKFHNW	0.0806	0.0649	0.2961	0.1111	0.1295	0.1558	
	@FHNW Technik	@fhnw _i4ds	@IT _FHNW	@iwi fhnw	@ic _fhnw	@dot FHNW	@ITHGK FHNW

Zentralitäts-Masse und Prestige-Masse

Eigenvector Centrality

Mit der Eigenvector Centrality Metrik kann der Einfluss bzw. die (theoretisch) mögliche Reichweite der einzelnen Portale verglichen werden. Die Zahlen zeigen nur den theoretisch grösstmöglichen Einfluss auf. Ob die Followers die Tweets auch wirklich lesen, kann natürlich nicht überprüft werden. Und die maximale Reichweite wird auch nur erreicht, wenn die Tweets des Portals von den Followers retweetet werden, was wir im Rahmen dieser Studie nicht untersucht haben, weil die dafür benötigte Datenmenge zu umfangreich gewesen wäre[10].

Twitter-Account	Eigenvector Centrality des Egonetzwerkes (Eigenvector Centrality Sum Change)
@FHNWTechnik	128.77519
@fhnw_i4ds	68.64615
@IT_FHNW	232.55462
@iwifhnw	700.71751
@ic_fhnw	16.26316
@ITHGKFHNW	115.58333
@dotFHNW	31.88235

Grundsätzlich lässt sich sagen, dass um so wichtiger die Nachfolger sind, umso bedeutsamer ist der Knoten (Portal) selbst.

[10] Es müssten dazu alle Tweets der Followers untersucht werden. Natürlich wäre es möglich den Publikations-Zeitraum der untersuchten Tweets einzuschränken oder nur einen Prozentanteil der Tweets über einen Zufallswert auszuwählen.

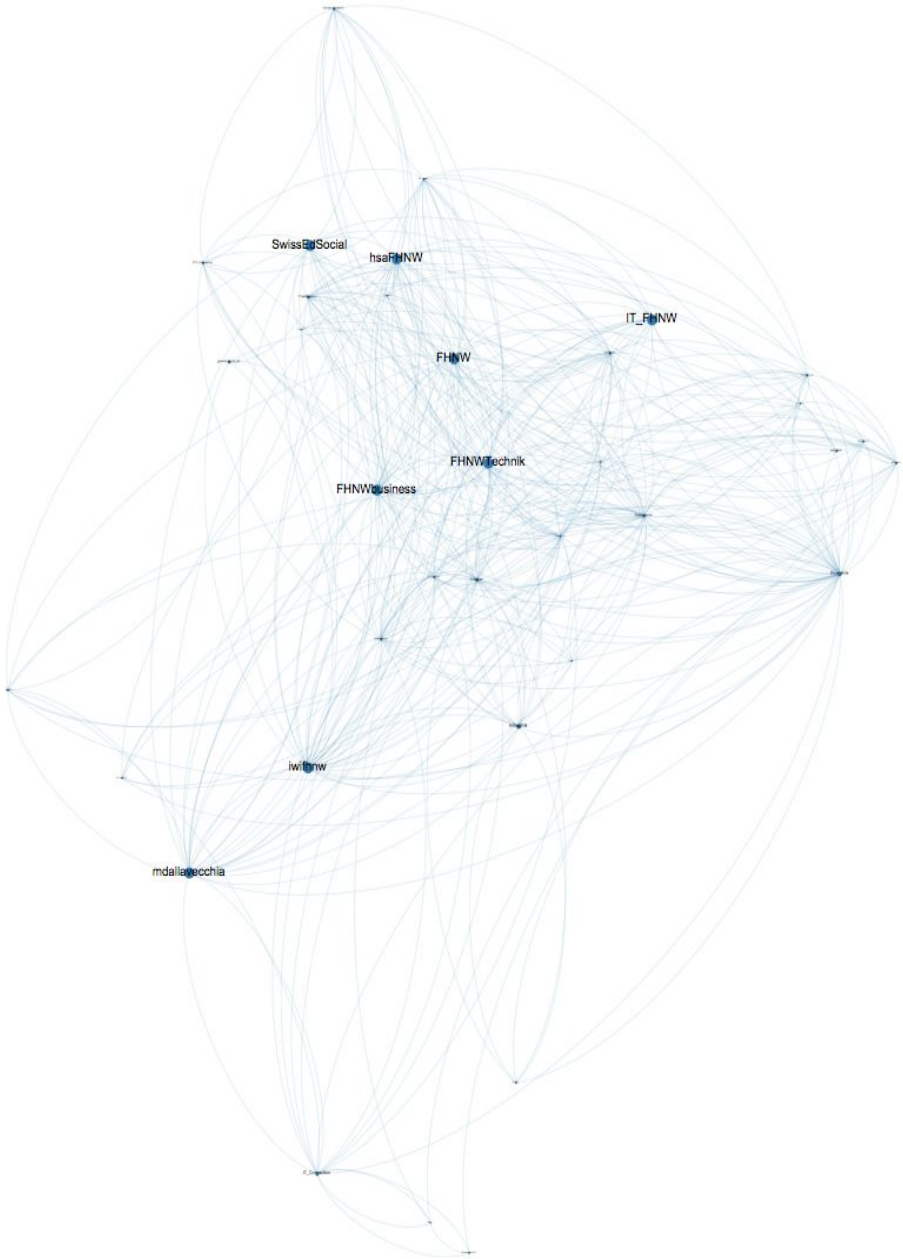
Clusters und Communities

Fragestellung:
“Wie stark sind die Followers eines bestimmten Accounts untereinander vernetzt (z.B. n-Clique oder k-Core)?”

Der Clustering Coefficient misst die Verbindungen zwischen Nachbarn, es ist ein Mass für Cliquenbildung. Die untenstehende Tabelle zeigt auf, in welchen Egonetzwerken die stärksten Verbindungen zwischen Nachbarn bestehen. Die Graph Density ist eine weitere Metrik, die Auskunft über die Vernetzung innerhalb eines Graphen gibt.

Twitter-Account	Anzahl Nodes im Ego-netzwerk	Anzahl Edges im Ego-netzwerk	Clustering Coefficient des Ego-netzwerks	Graph density	Anzahl “strongly connected compontents” des Egonetzwerkes	Nodes/ Components (durchschnittliche Grösse einer Component)
@FHNWTechnik	139	1319	0.489	0.069	29	4.79
@fhnw_i4ds	78	608	0.518	0.101	8	9.75
@IT_FHNW	241	2010	0.563	0.035	99	2.43
@iwifhnw	720	13625	0.437	0.026	155	4.65
@ic_fhnw	20	71	0.396	0.187	8	2.5
@ITHGKFHNW	121	650	0.434	0.045	65	1.86
@dotFHNW	39	106	0.329	0.089	16	2.44

Der Account @IT_FHNW weist den höchsten Clustering Coefficent auf gefolgt von @fhnw_i4ds. Das bedeutet, die Followers von diesen Twitteraccounts sind am stärksten miteinander verknüpft. Beim @fhnw_i4ds bilden 36 Knoten, also fast die Hälfte aller Nodes, ein 15-Core, und auch bei @IT_FHNW haben 110 Knoten diese Eigenschaft.



Die Abbildung zeigt die Knoten im Egonetzwerk vom Account @fhnw_i4ds auf. Sie bilden ein 15-Core.

Aktivität der untersuchten Accounts

Fragestellung:
“Wie aktiv sind die Accounts (Häufigkeit der Tweets)? Wie hoch ist der Relevanz der Tweets?”

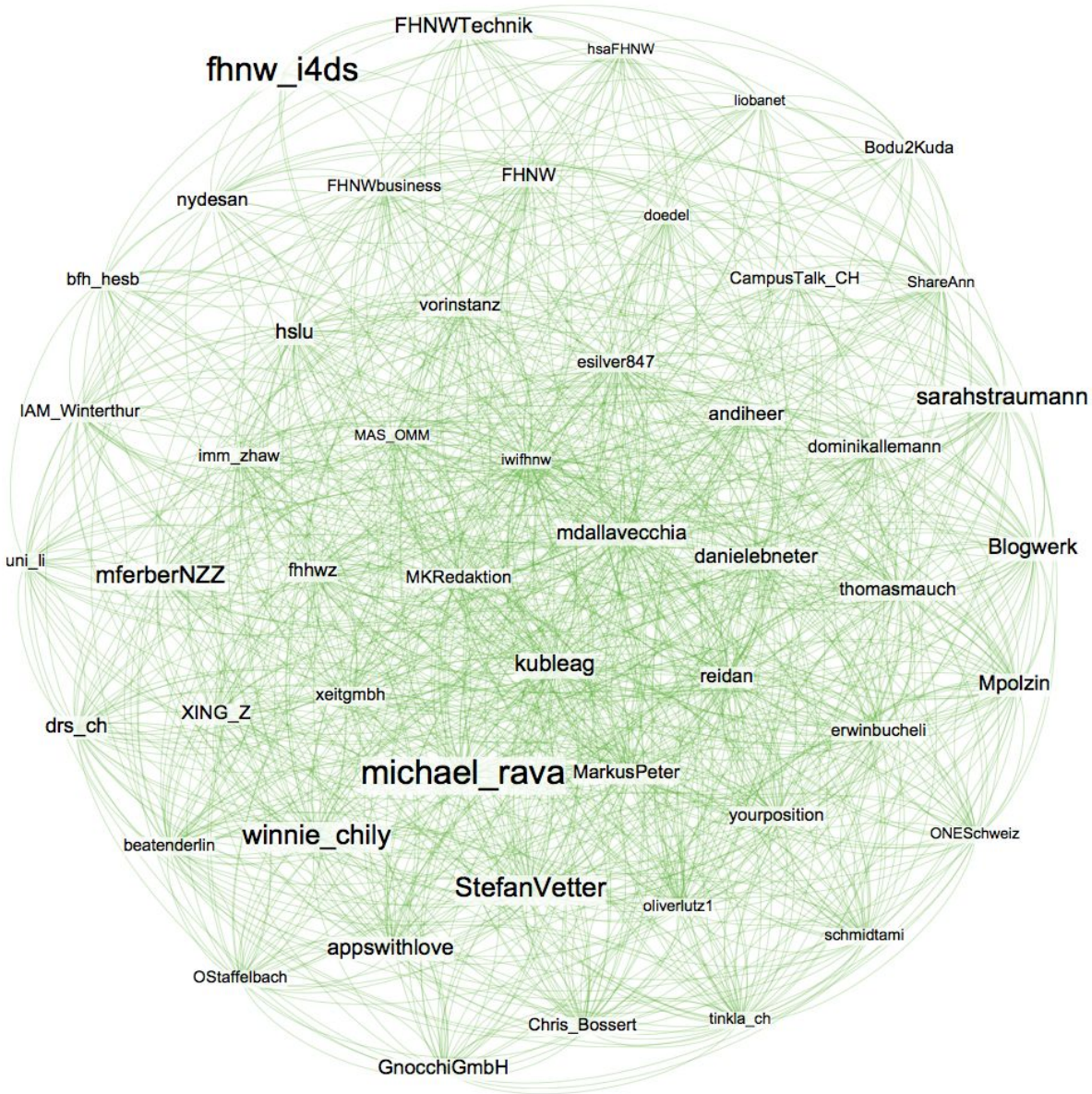
Die nachfolgende Tabelle gibt eine Übersicht der Aktivität der einzelnen Accounts. Dabei werden nicht nur die Anzahl Tweets verglichen, sondern diese auch in Relation zur Dauer, seit wann der Benutzer auf Twitter registriert ist, gestellt. Die daraus resultierenden Werte sind Durchschnittswerte. Es kann gut möglich sein, dass sich diese im Verlauf der Zeit verändert haben und, dass ein Benutzer innerhalb eines kurzen Zeitfensters sehr aktiv war. Um dies herauszufinden, hätten jedoch die einzelnen Tweets ebenfalls untersucht werden müssen, was nicht Teil unserer Projektarbeit war. Es kann aber sicher davon ausgegangen werden, dass viele Likes auf eine hohe Relevanz dieser Tweets für den User deuten.

Twitter-Account	Anzahl Tweets	Auf Twitter registriert seit	durch-schnittliche Anzahl Tweets pro Woche	durch-schnittliche Anzahl Tweets pro Tag	Anzahl Likes	Likes/ Tweets Ratio
@FHNWTechnik	55	2015-11-05 (36 Tage)	10.8	1.53	13	0.2364
@fhnw_i4ds	398	2015-01-26 (319 Tage)	8.7	1.25	69	0.1734
@IT_FHNW	419	2012-06-11 (1278 Tage)	2.3	0.33	32	0.0764
@iwifhnw	1574	2009-10-16 (2277 Tage)	4.8	0.69	263	0.1671
@ic_fhnw	16	2010-07-16 (1974 Tage)	0.06	0.008	0	0
@ITHGKFHNW	363	2011-05-27 (1659 Tage)	1.5	0.2	0	0
@dotFHNW	34	2010-09-28 (1900 Tage)	0.13	0.018	0	0

Vergleich verschiedener Rankings

Wie kann Einfluss gemessen werden? Es gibt verschiedene Metriken, die dafür eingesetzt werden können und je nach dem welche Werte verglichen werden unterscheiden sich die Rankings leicht voneinander. Abgesehen vom Followers/Following-Ratio ist die Tendenz jedoch immer die Selbe. Von den von uns untersuchten Accounts scheint @iwifhnw der einflussreichste zu sein.

Account	Anzahl Followers	Followers/ Following-Ratio	Anzahl einfluss-reiche Followers >=1000	Eigen-vector Centrality	Average Degree im Ego-netzwerk	Anzahl gelikte Tweets
@iwifhnw	710 (1.)	0.9233 (6.)	146 (2.)	700.72 (1.)	18.961 (1)	263 (1.)
@IT_FHNW	237 (2.)	4.6471 (3.)	178 (1.)	232.56 (2.)	8.340 (3.)	32 (3.)
@FHNWTechnik	135 (3.)	0.9375 (5.)	33 (3.)	128.78 (3.)	9.489 (2.)	13 (4.)
@ITHGKFHNW	120 (4.)	120 (1.)	6 (5.)	115.58 (4.)	5.372 (5.)	0
@fhnw_i4ds	65 (5.)	0.3988 (7.)	18 (4.)	68.65 (5.)	7.795 (4.)	69 (2.)
@dotFHNW	34 (6.)	2.4286 (4.)	3 (6.)	31.88 (6.)	3.029 (7.)	0
@ic_fhnw	19 (7.)	9.5 (2.)	2 (7.)	16.26 (7.)	3.550 (6.)	0



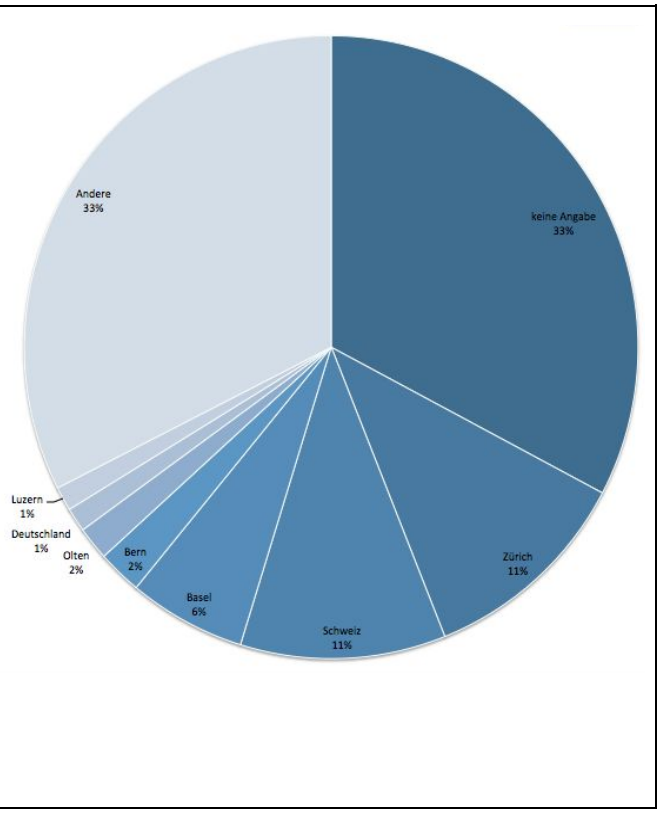
Die Abbildung zeigt das Egonetzwerk von @iwifhnw, wobei nur die Followers mit einem Mutual Degree Range von über 40 dargestellt wurden. Die Grösse der Tags visualisiert das Gewicht (Anzahl Follower) der jeweiligen Knoten. @iwifhnw scheint in verschiedener Hinsicht der einflussreichste von den von uns untersuchten Accounts zu sein.

Demografie der Followers

Location

Die nachfolgendende Tabelle gibt Auskunft über die häufigsten Ortsangaben der User unseres gesamten Netzwerkes. Dieses Feld kann vom Twitteruser frei erfasst werden. Die Angaben folgen keinem standardisierten Schema und auch die Korrektheit der Angaben wird in keiner Form überprüft.

Ortsangabe [11]	gesamtes Netzwerk, %	gesamtes Netzwerk, absolut
keine Angabe	32.75%	339
Zürich (24)	11.21%	116
Schweiz (15)	10.82%	112
Basel (15)	6.21%	64
Bern	2.33%	24
Olten	1.75%	18
Deutschland	1.26%	13
Luzern	1.26%	13
Andere[12]	32.45%	336



Die Followers kommen aus folgenden Ländern (alphabethisch):

Argentinien, Australien, Bangladesh, BVI (British Virgin Islands), Deutschland, Equador, Finnland, Frankreich, Indien, Irland, Italien, Kanada, Kasachstan, Korea, Lichtenstein, Luxemburg, Neuseeland, Österreich, Quatar, Schweiz, Singapur, Slowenien, Spanien, USA, Vereinigtes Königreich sowie Wales.

Weitere Bemerkungen zur Ortsangabe

Die Bezeichnungen der Locations variieren zwischen detaillierter Ortsangabe inkl. Adresse, Stadt, Region, Kanton, Land, Länderregion (z.B. D-A-CH) oder Kontinent. Einzelne User haben die PLZ oder GPS Koordinate angegeben, manche einen nicht reellen Ort (z.B. “my little space of paradise”, “World Cityzen”, “mal hier mal da”) oder die Angaben sind extrem vage definiert (“Orion-Cygnus Arm, Milky Way”, “Around the World” oder “Space”). Eine weitere Herausforderung ist, dass manche mehrere Ortsangaben erfasst haben. Dazu kommen die unterschiedlichen Schreibweisen, z.B für Schweiz gab es in unserer Benutzergruppe 15

[11] Angabe in Klammern beschreibt die Anzahl unterschiedlicher Schreibweisen
[12] weniger als 1% pro Location

verschiedene Schreibweisen und für Zürich sogar 24! Eine weitere Herausforderung liegt darin, dass einzelne Benutzer bei der Ortsangaben ein anderes Alphabet verwenden (wahrscheinlich kyrillisch).

Ortsangaben @fhnw_i4ds, nach Ländern gruppiert
Das Egonetzwerk des Accounts @fhnw_i4ds scheint eine deutlich internationalere Verteilung der Followers aufzuweisen, wie folgende Tabelle aufzeigt:

Ortsangabe Land	Beinhaltete Orte[13]	@fhnw_i4ds, %	@fhnw_i4ds, absolut[14]
keine Angabe	-	20.51%	16
Schweiz	Zürich (10), Brugg (6.5), Basel (2.5), Olten (2.5), Bern (1.5), Aarau, Luzern, Genf, Nordwestschweiz	48.72%	38
USA	Jamestown NY (2), Hawaii, San Francisco, Austin TX, TN (Tennesee?)	7.69%	6
Deutschland	Ingolstadt, Münster, Dresden, Hamburg	6.41%	5
UK	London, Stevenage	2.56%	2
Australien	Hobart	1.28%	1
Finland	-	1.28%	1
Frankreich	-	1.28%	1
Luxenburg	Luxenburg	1.28%	1
Österreich	Salzburg	1.28%	1
Europa	-	1.28%	1
Anderes	Space, Milky Way, no place	3.85%	3
URL	-	2.56%	2

[13] in absteigender Reihenfolge, absolute Zahlen in Klammern.

[14] es hat bei den absoluten Angaben Fliesskommazahlen, da bei mehrere Ortsangaben im Locationsattribut diese Anteilsmässig berechnet wurden

Sprache

Die nachfolgendende Tabelle gibt Auskunft über die häufigsten Sprachangaben der User unseres gesamten Netzwerkes sowie der einzelnen Teilnetzwerke. Die Sprachangabe ist etwas standardisierter als die Ortsangabe erfasst, sie hat aber nicht immer eine Korrelation mit der Herkunft des Benutzers. Twitter empfiehlt dem Benutzer die Sprache zu deklarieren, in welcher er in der Regel seine Tweets schreibt (was ja nicht immer die Muttersprache des Benutzers ist).

	DE	EN[15]	FR	IT	ES	RU	andere
gesamtes Netzwerk	66.96%	29.18%	1.64%	0.48%	0.39%	0.39%	0.98%
@FHNWTechnik	71.22%	25.9%	2.16%	-	-	0.72%	-
@fhnw_i4ds	47.44%	48.72%	2.56%	-	-	1.28%	-
@IT_FHNW	68.05%	28.63%	1.64%	0.41%	-	0.83%	0.83%
@iwifhnw	68.47%	27.08%	1.81%	0.56%	0.56%	0.28%	1.24%
@ic_fhnw	70%	25%	-	-	-	5%	-
@ITHGKFHNW	70.25%	28.1%	0.83%	0.83%	-	-	-
@dotFHNW	57.14%	42.86%	-	-	-	-	-

Wir sehen in der Tabelle gewisse Unterschiede zwischen den einzelnen Teilnetzwerken. Besonders beim Teilnetzwerk von @fhnw_i4ds sowie @dotFHNW hat es deutlich mehr User als bei den andern Teilnetzwerken, welche Englisch als Sprache erfasst haben; bei @fhnw_i4ds ist Englisch sogar die häufigste Sprache.

[15] EN sowie EN-GB kummuliert

Fazit

Selektion des zu untersuchenden Netzwerkes

Wie bereits in der Analyse beschrieben mussten wir im Laufe des Projektes den Scope mehrmals ändern. Unsere ursprüngliche Idee war verschiedene Schweizer Newsportale zu untersuchen. Daraus ist jedoch ein viel zu grosses Netzwerk entstanden (über 800'000 Nodes). Nach der Reduktion der Daten haben wir realisiert, dass die Beziehungen unter den einzelnen Nodes grösstenteils verloren gegangen waren, so dass nicht mehr all zu viele interessante Netzwerkanalysen gemacht werden konnten. Nach diesem ersten "Fehlversuch" haben wir dann eine andere Domäne gesucht, auf die wir unsere initiale Idee und Fragestellung übertragen konnten.

Weil wir den Code des Fetchers mit kleinen Ausnahmen recht generisch gehalten hatten, konnte dieser nach unseren "Fehlversuchen" relativ rasch für die neuen Analysegebiete adaptiert werden.

Wir haben aus den Fehlversuchen gelernt, dass sich die Beziehungen auf Twitter teilweise stark von denen in der realen Welt unterscheiden und nicht alle natürlichen Beziehungen auf Twitter abgebildet sind. Es lohnt sich auf jeden Fall Stichproben der Daten zu erheben, bevor viel Zeit ins Fetchen der Daten investiert wird.
Idealerweise sollte immer das ganze Netzwerk gefetcht werden können so dass nicht mit reduzierten, und somit verfälschten Daten, gearbeitet werden muss. Falls dennoch die Daten reduziert werden müssen, sollte dies nicht einfach zufällig geschehen. Die Kriterien für die Filterung sollten gut gewählt werden, damit die Aussagekraft der Daten erhalten bleibt.

Demografische Filterung

Eine weitere "Lesson learned" ist, dass die demografische Filterung der Personendaten von Twitter Benutzern eher herausfordernd ist. Die Daten sind oft nicht vorhanden oder nicht standardisiert erfasst. Es ist schwer Aussagen über die Demografie der Followers zu machen, da grundsätzlich nur Sprache, Ort und Bio (Beschreibung) vom User vorhanden sind. Für weitere demografische Indizien zur Person wie zum Beispiel Alter, Geschlecht oder Beruf müssten versucht werden, diese Attribute aus der Bio herauszufiltern. Die Bio ist jedoch ein vom Benutzer frei erfasster Text und es ist sehr unterschiedlich im Inhalt. Zudem haben sowieso nur 22.6% der Benutzer aus dem Newsportal Netzwerk bzw. 68.7% der Personen aus dem FHNW Informatik Netzwerk überhaupt eine Bio erfasst.

Bei der Sprache wird dem Benutzer von Twitter empfohlen die Sprache anzugeben, in welcher die Tweets hauptsächlich verfasst werden, was nicht in jedem Fall die Muttersprache des Users ist.

Dasselbe gilt für den Wohnort, dieser kann vom User frei gewählt werden und wird als Textstring hinterlegt. Die Angaben variieren stark was den Detailgrad der Location betrifft, wie wir bereits im entsprechenden Abschnitt der Analyse festgehalten haben. Da kein standardisiertes Schema wie z.B. Country Code oder Kantonsabkürzungen besteht, müssen die Inhalte zuerst aufbereitet werden, bevor eine Analyse möglich ist.

Sentimentanalyse von Tweets

Unsere erweiterte Fragestellung beinhaltete auch die Untersuchung von Tweets. Obschon gerade die Sentimentanalyse ein sehr spannendes Thema wäre und Twitter sogar unterstützende Funktionen dazu in der API anbietet, hat sich schnell herausgestellt, dass dies ein zu umfangreiches Unterfangen wäre und die Erkenntnisse daraus nicht unsere Kernfrage beantworten würden.

Messung von Einfluss

Unsere Fragestellungen waren teilweise etwas subjektiv gestellt. Einfluss beispielsweise kann auf unterschiedliche Arten gemessen werden, z.B. Anzahl Followers (Prestige Indegree), Anzahl Follower der Followers (Prestige Proximity oder Eigenvector Centrality), Vorkommen von besonders bedeutenden Followers (Filterung der Followers z.B. nach deren Anzahl Followers, Anzahl Tweets), Followers/Following-Ratio, …). Je nach gewählter Metrik führt die Analyse zu unterschiedlichen Resultaten. Bei manchen Metriken sind die Unterschiede minimal, bei andern bedeutend (z.B. Followers/Following Ratio).

Wenn eine subjektive Notation der Resultate vermieden werden will, ist es wichtig die Fragestellung messbar und sachlich-neutral zu formulieren und am besten von Anfang an Metriken zu definieren, mit welchen die Resultate gemessen werden müssen. Ansonsten besteht die Möglichkeit, dass die Analyse durch die Wahl der Metriken gefärbt werden könnte, in dem gezielt eine Metrik eingesetzt wird, welche die gewünschte Aussage unterstreicht.

Datenmenge

Einige unserer initialen Fragestellungen waren zudem nicht beantwortbar, weil eine zu grosse Datenmenge dazu nötig gewesen wäre, beispielsweise die Analyse der Tweets und Retweets. Wir hätten dazu nicht nur mit dem Rate-Limit von Twitter gekämpft sondern wären auch an die Grenzen von Gephi gestossen.
Auch unser erster Versuch mit der Analyse der Newsportale hat uns gezeigt, dass eine Reduktion der Datenmenge oft mit Verlusten der Aussagekraft verbunden ist und die Art der Datenreduktion gut durchdacht sein muss. Ob die Daten nach Zufallswert reduziert werden oder nur die aktuellsten Einträge analysiert werden hat auf jeden Fall einen Einfluss auf die Resultate der Analyse.

Gephi

Gephi ist nicht immer ganz intuitiv in der Anwendung und braucht einiges an Geduld und Routine. Besonders mühsaeiten, wenn man mal herausgefunden hat wie diese erzeugt werden können. m ist, dass gewisse Werte nur über Sliders eingegeben werden können und damit eine präzise Eingabe von Werten (z.B. um einen Range zu definieren) manchmal fast unmöglich ist. Einige Analysen waren für uns einfacher mit Java-Code oder mit Excel auszuführen, das uns diese Tools vertrauter sind. Dafür bietet Gephi differenziertere, eindrucksvollere Darstellungsmöglichk