Social Network Mining

David Theil B.Sc.

Seminararbeit

eingereicht am

Fachhochschul-Masterstudiengang

Information Engineering und Management

im Juli 2012

Inhaltsverzeichnis

1. Einleitung

1.1. Abstract

This article is about social network analytics. In the first part the motivation and the reasons for social network analytics will be shown. After that there will be a definition for social networks, social network services and multi user virtual environments and the elements of social networks will be discussed. The next part will explain some basic methods of static and dynamic structural social network analysis. After that, other methods for analyzing a social network like link prediction, node classification, and community detection will be described. Furthermore it will be explained how multimedia mining in social networks works and how to visualize social network graphs.

1.2. Kurzfassung

In dieser Arbeit wird ein Überblick über Analyse sozialer Netzwerk geben. Zunächst wird die Motivation und der Grund für die Analyse soziale Netzwerk erklärt. Danach werden wichtige Begriffe wie soziales Netzwerk und soziale Netzwerkservices definiert und die wichtigsten Bestandteile von sozialen Netzwerken beschrieben. Nach diesen Grundlagen werden einige Methoden zur statischen und dynamischen strukturellen Netzwerkanalyse beschrieben. Danach wird auf andere Analyseverfahren eingegangen wie Community Detection, Knotenklassifizierung, und Linkschlussfolgerung. Des Weiteren wird auf Multimedia Mining in sozialen Netzwerken und Visualisierung von sozialen Netzwerkgraphen eingegangen.

1.3. Motivation

Social Network Service Sites, wie beispielsweise Facebook, MySpace oder LinkedIn haben in den letzen Jahren einen wahren Boom erlebt. Mit dem Auftreten von sozialen Netzwerkseiten im Internet und der steigenden Beliebtheit dieser bei den Internetnutzern/Internetnutzerinnen ist nun auch die Analyse von sozialen Netzwerken wieder mehr von Bedeutung [SCOT 2010]. Da die Kosten für internetfähige Geräte wie Smartphones und Netbooks in letzter Zeit stark gefallen sind, ist die Zahl der Mitglieder dieser sozialen Netzwerke um ein Vielfaches angewachsen [AGGA 2011].
Aus soziologischer Sicht werden soziale Netzwerke bereits seit den 1930er Jahren erforscht [SCOT 2010]. Die soziale Netzwerkanalyse profitiert von dem Boom der Social Network Service Sites, da nun erstmals riesige Mengen an Daten in elektronischer Form für die

Forschung zur Verfügung stehen. Durch diese Daten können nun erstmals Theorien über soziales Verhalten von Personen oder Personengruppen sowie Beziehungen zwischen Personen einfach überprüft werden und neue Forschungen und Auswertungen erfolgen.

1.4. Zielsetzungen

In dieser Seminararbeit wird das Thema Social Network Mining bearbeitet. Diese Arbeit soll den momentanen Stand der sozialen Netzwerkanalyse aufzeigen, in welchen Bereichen geforscht wird und welche Ergebnisse und Forschungsprojekte bestehen. Es soll dem Leser/der Leserin ein grundsätzlicher Eindruck über die Materie der sozialen Netzwerkanalyse gegeben werden. Es sollen auch elementare Eigenschaften von sozialen Netzwerken beschrieben und grundlegende Algorithmen zur Analyse erklärt werden. Des Weiteren sollen technische Methoden für Social Network Mining recherchiert und beschrieben werden. Es soll dargestellt werden, welche Auswertungen bereits möglich sind und eventuell schon durchgeführt werden.

2. Soziale Netzwerke

In diesem Kapitel werden wichtige Begriffe der Analyse sozialer Netzwerk definiert sowie der Aufbau sozialer Netzwerke erläutert und die grundlegenden Bestandteile von sozialen Netzwerken beschrieben.

2.1. Definition soziale Netzwerke

Es gibt mehrere Definitionen von sozialen Netzwerken in der Literatur. Charu C. Aggarwal definiert ein soziales Netzwerk als ein Netzwerk aus Interaktionen oder Beziehungen. Die Knoten in diesem Netzwerk repräsentieren Akteure die Kanten sind Beziehungen oder Interaktionen zwischen diesen Akteuren [AGGA 2011].
Ganz ähnlich definierten S. Wasserman und K Faust bereits 1994 ein soziales Netzwerk als ein System mit einer Menge an sozialen Akteuren und einer Sammlung von sozialen Beziehungen, welche spezifizieren wie diese Akteure zueinander in Beziehung stehen [WASS et al. 1994].

2.2. Definition Social Network-Services (SNS)

Ein SNS Social Network Service oder auch Online Social Network genannt, ist eine multifunktionale Onlineplattform, welche es ermöglicht persönliche Inhalte zu erstellen, Fotos und Videos zu teilen, Nachrichten untereinander auszutauschen, zu Bloggen und Content anderer Users zu kommentieren. Die Hauptfunktion jedoch zeigt an, wer mit wem befreundet ist und es Benutzern ermöglicht freundschaftliche Kontakte zu pflegen [ROSE 2010].
Das Forschungsinteresse an diesen Sozialen Netzwerk Services ist in der letzten Zeit stark gestiegen. Es wird an Studien über das soziale online Kapital und soziale Ressourcen, [ELLI et al. 2007], [LACK et al. 2009], [STEF 2010] geforscht. Soziales online Kapital und soziale Ressourcen können beispielsweise wichtige Kontakte in der Freundesliste sein, welche den passenden Job haben den der User gerade sucht.

3. Analyse von sozialen Netzwerken

Die Erforschung und Analyse sozialer Netzwerke, hat sich in den letzten Jahren parallel zur computergestützten Kommunikation entwickelt. Die computerunterstützte Datenverarbeitung und das Internet erlauben es heute Netzwerkforschern, soziale Phänomene mit neuen Methoden zu untersuchen. Zusätzliche Fortschritte in der strukturellen Analyse und Visualisierung von sozialen online Netzwerken wurden ebenfalls erzielt.
Social Network Service Sites wie beispielsweise Facebook, Blogs, social Media und andere digitale soziale Netzwerke beinhalten bereits riesige Datenmengen. Die nun zur Verfügung stehenden Daten erbringen einen neuen wichtigen Beitrag in der Netzwerkforschung. Durch diese Daten ist es beispielsweise möglich, große Datenströme über lange Zeiträume zwischen den Knoten in den Netzwerken zu analysieren, einzigartige Beziehungsdaten über eine Vielzahl von Knoten zu erforschen. Dynamische Daten erlauben eine längerfristige Analyse und Animation von sozialen Netzwerken über die Zeit und ermöglicht es beispielsweise die Entwicklung dieser Netzwerke zu dokumentieren, darzustellen und auszuwerten [ROSE et al. 2010].

3.1. Aufbau sozialer Netzwerke

Soziale Netzwerke können als Graph dargestellt werden. Das soziale Netzwerk ist gleichbedeutend mit einen Graph $G = (N, E)$ bestehend aus einer Menge an Knoten N oft auch V und einer Menge an Kanten E [AGGA 2011].
Dieser Graph kann gerichtet oder auch nicht gerichtet sein. Beispielsweise wird eine Freundschaftsbeziehung zwischen zwei Personen (Knoten n_i und n_j in der Menge N) mit einer nicht gerichteten Kante (Kante $e=(n_i, n_j)$ der Menge E) dargestellt [AGGA 2011]. Ein Telefonanruf oder eine E-Mail zwischen den zwei Knoten wird jedoch als gerichtete Kante dargestellt, da es sich hierbei um eine Aktion vom Sender zum Empfänger handelt und somit eine Richtung aufweist. Würde diese Kante nicht gerichtet dargestellt ginge Information verloren.

3.1.1. Uniparte und Multiparte Graphen

Es gibt uniparte und multiparte Graphen. Uniparte Graphen sind homogen und bestehen somit aus nur einen Knotentyp, Multiparte Graphen sind hingegen inhomogen und bestehen aus unterschiedlichen Knotentypen. [AGGA 2011]

Als Beispiel für einen gerichteten uniparten Graphen siehe Abbildung 1; Einen multiparten nicht gerichteten Graphen zeigt Abbildung 2.

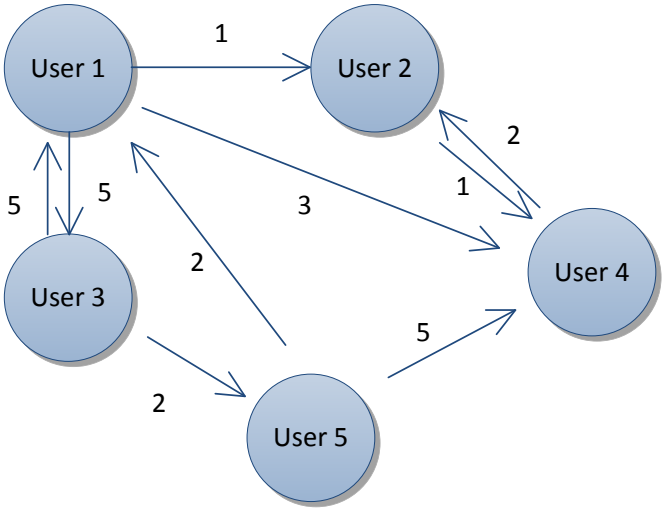

Abb. 1 gerichteter uniparter Graph

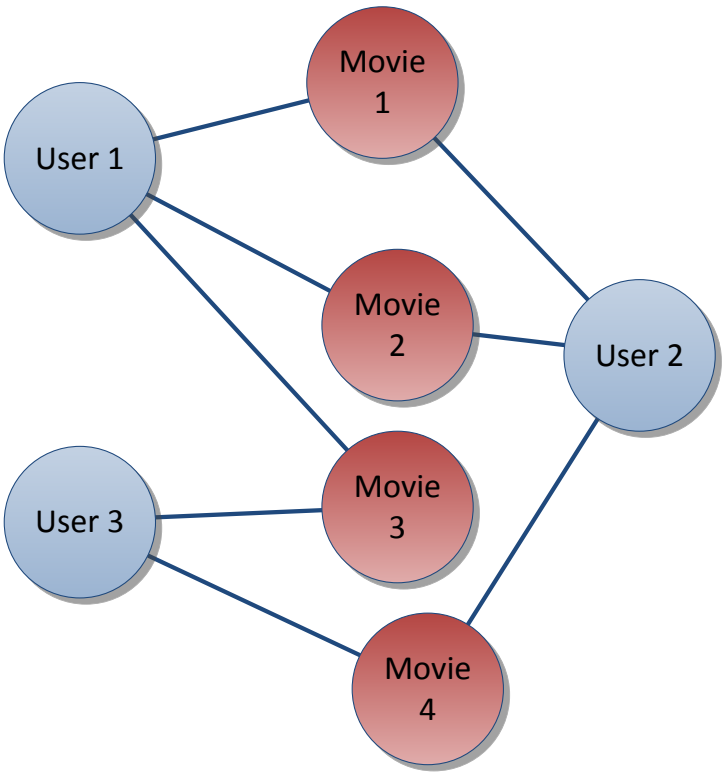

Abb. 2 nicht gerichteter multiparter Graph mit 2 Knotentypen Users und Movies [AGGA 2011; Abb. 2.1]

3.1.2. Kantengewichte

Graphen mit mehreren Kanten zwischen zwei Knoten können in zwei unterschiedliche Arten dargestellt werden. Einerseits mit Multiedges, also mit mehreren Kanten zwischen zwei Knoten oder mit Gewichten über einer Kante.
Wenn mehrere unterschiedliche Interaktionen zwischen zwei Knoten vorkommen, wird häufig die Variante mit mehreren Kanten zwischen den Knoten gewählt. Wenn eine Aktion mehrmals abgehalten wird, wird die Variante mit den gewichteten Kanten verwendet [MCGL et al. 2011].
Bei gerichteten Graphen kann es vorkommen, dass mehrere Kanten zwischen zwei Knoten notwendig sind. Beispielsweise können mehrere Emails zwischen zwei Personen mittels mehrerer gerichteter Kanten dargestellt werden. Diese Methode kann jedoch sehr schnell zu unüberschaubar werden. In diesem Fall kann eine Darstellung mit gewichteten Kanten eingesetzt werden [AGGA 2011].

3.2. Strukturelle Analyse von sozialen Netzwerken

Die folgenden Abschnitte sind den verschiedenen Kennzahlen sozialer Netzwerke gewidmet.
Diese Kennzahlen geben eine Antwort über verschiedene Netzwerkgrößen. Sie können Regelmäßigkeiten und Unregelmäßigkeiten von sozialen Netzwerken aufzeigen. Anomalien dieser Verhaltensmuster geben Hinweise auf Missbrauch von Funktionen sozialer Netze wie beispielsweise Link-Spamming, oder das Fälschen von positiven Bewertungen in e-Auctions Systemen, oder abnorme Subgruppen in Social Network Seites wie Facebook, Yahoo-360 oder LinkedIn [LESK et al. 2005]. Doch die Analyse von sozialen Netzwerkgraphen kann nicht nur Missbrauch aufdecken, sie kann auch hilfreich beim Auffinden von nützlichen Netzwerkeigenschaften sein, wie beispielsweise das Identifizieren von einflussreichen Benutzern [BORO et al. 2005], [CHAK et al. 1999], [KUMA et al. 2006].

3.2.1. Statische Analyse von sozialen Netzwerken

Obwohl alle realen sozialen Netzwerke sich über die Zeit entwickeln und somit der Zeitbezug bei der Analyse wichtig ist, gibt es Analysemethoden die diesen Aspekt außer Acht lassen und nur Werte für einen einzigen Moment, also einen Snapshot des Graphen erfassen. Diese Analysemethoden befassen sich mit den statischen Eigenschaften von sozialen Netzwerken [MCGL et al 2011].

3.2.1.1. Komponentenverteilung

Eine weitere wichtige Eigenschaft von Graphen ist die Komponentenverteilung. Eine Komponente besteht aus einer Menge an Knoten und Kanten die untereinander so verbunden sind, dass man von jedem Knoten in der Komponente jeden anderen erreichen kann. In realen Graphen kann man beobachten, dass sich die Knoten über die Zeit zu einer einzigen großen Komponente vereinigen [MCGL et al. 2011].

3.2.1.2. Durchmesser und effektiver Durchmesser

Zur Beobachtung und zur Bewertung von Veränderungen über die Zeit im Graphen und vor allem bei den einzelnen Komponenten im Graphen, kann der Durchmesser als Messgröße herangezogen werden. Der Durchmesser einer Komponente oder eines Subgraphen wird bestimmt indem man die maximale Distanz zwischen zwei Knoten im Graphen berechnet. Die Distanz ist dabei die Anzahl an Hops auf dem kürzesten Pfad von Knoten A zum Knoten B [MCGL et al. 2011].
In der sozialen Netzwerk Analyse kann der kleinste Durchmesser leicht bestimmt werden. In realen Netzwerken kann das „Small-World-Phenomenon“ oder „Six degrees of seperation“ beobachtet werden, bei dem der kleinste Durchmesser meist zwischen 6 und 7 liegt [MCGL et al. 2011]. Dabei wird die maximale Entfernung (in Hops) von zwei beliebigen Knoten im Graphen berechnet, wobei immer nur die kleinste Anzahl an Hops aller möglicher Pfade zwischen den beiden Knoten zur Berechnung herangezogen wird. Bei dieser Berechnung wird die Richtung der Kante meist nicht berücksichtigt.
Da bei dieser Messung nur das Maximum der kürzesten Wege durch den Graphen zur Kalkulation verwendet wird, bleiben viele sehr lange Pfade durch den Graph unberücksichtigt und scheinen somit nicht in der Kennzahl auf. Um diese Schwachstelle zu umgehen benutzt man auch oft den effektiven Durchmesser (effective diameter). Er gibt die minimale Anzahl an Hops an, bei dem eine gewisse Menge (normalerweise 90%) aller verbunden Knoten erreicht werden kann [MCGL et al. 2011].

3.2.1.3. Community Structure (gesellschaftliche Struktur)

Soziale Graphen weisen eine modulare Struktur vor, indem sie Gruppen und sogar Gruppen mit Untergruppen bilden. Knoten einer Gemeinschaft verbinden sich mehr untereinander und weniger mit Knoten anderer Gemeinschaft [MCGL et al. 2011]. Die Anzahl an Gruppen in einem sozialen Netzwerk ist somit eine wichtige Kennzahl in der sozialen Netzwerkanalyse.

3.2.2. Potenzgesetze

In sozialen Netzwerken kann man bei unterschiedlichen Kennzahlen beobachten, dass sich diese Zahlen sich nach dem Potenzgesetz verändern. Die folgenden Abschnitte geben einen Auszug aus diesen Kennzahlen.

3.2.2.1. Verteilungsgradpotenzgesetz

Der Verteilungsgrad in einem Graphen unterliegt dem Potenzgesetz in der Form $f(d) \alpha d^{-\alpha}$ wobei der Exponent $\alpha > 0$ ist und $f(d)$ ist die Teilmenge an Knoten der Menge N mit dem Grad d [MCGL et al. 2011]. Das Potenzgesetz in Bezug auf den Verteilungsgrad lässt vermuten, dass es viele zueinander nahestehende Knoten und nur einige wenige zueinander weitverteilte Knoten in wirklichen Graphen gibt [MCGL et al. 2011].

3.2.2.2. Dreiecks-Potenzgesetz (Triangle Power Law TPL)

Das Dreiecks Potenzgesetz besagt, dass die Anzahl von Dreiecken und die Anzahl von Knoten, welche die Eckpunkte dieser Dreiecke bilden, dem Potenzgesetz folgt [MCGL et al. 2011]. Somit gibt es viele Knoten die an nur wenigen Dreiecken mit benachbarten Knoten teilnehmen, aber wenig Knoten die an vielen Dreiecken im Graphen teilnehmen[MCGL et al. 2011].

3.2.2.3. Eigenwert-Potenzgesetz (Eigenvalue Power Low EPL)

Das Eigenwert Potenzgesetzt geht aus einer Konsequenz des Verteilungsgrad Potenzgesetzes hervor und bedeutet das die Eigenwerte der größten Entitäten im Graphen ebenfalls nach dem Potenzgesetz verteilt sind. Der Eigenwert wird berechnet, indem der Graph in einer Adjazenzmatrix abgebildet wird und der Eigenvektor mit dem Eigenwert für jede Spalte berechnet wird. Er gibt eine Aussage über die Zentralität eines Knotens im Vergleich zu den anderen Knoten. Betrachtet man den Eigenwert der Knoten über einen Zeitraum hinweg so wird beobachtet, dass der Eigenwert $\lambda_1(t)$ und die Anzahl der Kanten $E(t)$ miteinander den Potenzgesetzt folgen [MCGL et al. 2011]:

$$\lambda_1(t) \alpha E(t)^{\alpha}$$

3.2.2.4. Das Gewichtspotenzgesetz (Weight Power Law WPL)

Die zum Zeitpunkt t im Netzwerk befindlichen Knoten $E(t)$ sind mit einer Menge von Kanten mit dem Summengewicht von $W(t)$ verbunden. Hat sich nach einer gewissen Zeit die Anzahl der Knoten im Netzwerk verdoppelt, so ist die Summe aller Gewichte aller Kanten zwischen diesen Knoten um vieles größer als das Doppelte. Dies kann man mit der Formel für das WPL berechnen [MCGL et al. 2011]:

$$W(t) = E(t)^{w}$$

Dabei ist w der Gewichtsexponent und liegt meist zwischen 1.01 und 1.5 in realen Graphen [MCGL et al. 2011].

3.2.3. Dynamische Analyse von sozialen Netzwerken

Soziale Netzwerke entwickeln sich über die Zeit hinweg. Aus diesem Grund ist es wichtig die dynamischen Eigenschaften sozialer Netzwerke ebenfalls zu analysieren. Dies geschieht meist durch die Analyse einer Serie von statischen Snapshots dieser Netzwerke, bei der man die einzelnen Zeitpunkte in der Entwicklung des Netzwerks miteinander vergleicht [MCGL et al. 2011].
Die folgenden Abschnitte befassen sich mit Methoden der dynamischen Analyse von sozialen Netzwerken.

3.2.3.1. Gelierpunkt und Schrumpfender Durchmesser

In vielen realen Graphen, gibt es einen bestimmten Zeitpunkt, in dem der Durchmesser seinen Höhepunkt erreicht und wieder kleiner wird. Bevor dieser Punkt der Sättigung erreicht ist, besteht das Netzwerk aus einer lockeren Verbindung aus kleinen, losen Komponenten. Sobald jedoch eine bestimmte Sättigung erreicht wird („Gelling Point“) bildet sich eine riesige stark verbunden Komponente, „Giant Connected Component“ (GCC), welche dazu führt das der Durchmesser schrumpft oder sich stabilisiert und fortan nur mehr die GCC immer weiter wächst und sich mit den neuen Knoten verbinden. [MCGL et al. 2011]
In [LESK et al. 2005] zeigen die Autoren Leskovec et al., dass der Durchmesser eines realen sozialen Netzwerks nicht nur sehr klein ist, sondern auch schrumpft und sich dann über die Zeit hinweg stabilisiert. Das Netzwerk besteht aus vielen kleinen nichtverbundenen Komponenten, welche sich zu einer großen Komponente im Graphen vereinigen, indem sich neue Knoten als Verbindungen zwischen diesen Komponenten hinzufügen. Ab diesem Zeitpunkt beginnt der Durchmesser zu schrumpfen, bis er sich irgendwann eingependelt hat [MCGL et al 2011], [LESK et al. 2005].

3.2.3.2. Verdichtungspotenzgesetz

Das Verdichtungspotenzgesetz sagt aus, dass bei einer Verdopplung der Knoten im Netzwerk über die Zeit t, die Anzahl der Kanten zwischen den Knoten um mehr als das Doppelte ansteigt. Untersuchungen von Leskovence et al. ergaben, dass sich der Exponent der Vervielfachung zwischen 1,03 und 1,7 abbildet [LESK et al 2005], [MCGL et al. 2011].

3.3. Community Detection

Eine der wichtigsten und zugleich komplexesten Analysen von sozialen Netzwerken ist das Auffinden von Communities (Gruppen) innerhalb sozialer Netze. Die Community Detection ist mit dem Clustering von Graphen verwandt und versucht Regionen im Graphen mit einer höheren Dichte/höheren Anzahl an Vernetzungen zu finden [AGGA 2011]. Das Erkunden von Gruppen kann zum Beispiel dahingehend ausgenutzt werden, um Personen mit ähnlichen Interessen zu finden und somit gezielte personalisierte Werbung zu machen. Die Identifikation von einflussreichen Knoten, oder Subcommunities innerhalb von Bordercommunities können für virales Marketing genutzt werden [DOMI et al. 2001], [KEMP et al. 2003], [LESK et al. 2007].
Des Weiteren kann die Community Detection Aufschluss über verschiedene komplexe Beziehungen in Netzwerken geben und somit in den verschiedensten Bereichen wie Biologie, Ökonomie, Soziologie, Technologie, uvm. helfen, neue Zusammenhänge zu erkennen[PART et al. 2011]. Außerdem bietet sie Einsicht in die Strukturen, Eigenschaften und Verhaltensweisen einzelner Gruppen in sozialen Netzwerken [BARA et al. 2003].
Community Detection wird gleichermaßen bei statischen, als auch bei dynamischen Netzwerkanalysen eingesetzt.

3.3.1. Methoden und Algorithmen zur Community Detection

Im Allgemeinen kann eine Gruppe in einem Graphen als eine Ansammlung von Knoten mit einer höheren Anzahl an Verknüpfungen (verbindenden Kanten) untereinander als mit dem Rest des Graphen definiert werden.
Bei Algorithmen zur Community Detection kann es große Unterschiede im Laufzeitverhalten und anderen Eigenschaften geben. Eine dieser Eigenschaften ist, ob diese Algorithmen eine spezifische Qualitätsmetrik optimieren oder nicht. Mit der Qualitätsmetrik wird die Eigenschaft definiert, die alle Mitglieder dieser Gruppe aufweisen müssen, um der Gruppe anzugehören. Algorithmen wie der Kernighan-Lin Algorithmus und der Flow-Based Postprocessing Algorithmus versuchen die spezifische Qualitätsmetrik zu optimieren, während der Markov Clustering Algorithmus und das Clustering via Shingling das nicht tun. Eine weitere Eigenschaft die Algorithmen zu unterscheiden ist, ob sie dem User die Möglichkeit bieten die Anzahl der Communities, in die das Netzwerk eingeteilt wird, zu bestimmen. [PART et al. 2011]
Um die Güte der bestimmten Cluster des Graphen zu bestimmen werden sogenannte Qualitätsfunktionen angewendet. Die wichtigsten sind die *Normalized Cut Funktion (Ncut(S))* die *Conductance Funktion (Conductance(S))* und die *Modulartiy (Q).*
Es gibt Methoden und Algorithmen zur Community Detection, die sich mehr für die Analyse von statischen oder mehr für die Analyse von dynamischen Netzwerken eignen.
Die klassischen, wie beispielsweise, Community Detection via Shingling, Flow-Based Post-Processing, Local Graph Clustering, Recently, Regularized und Multilevel Markov Clustering, Multilevel Graph Partitioning, Girvan und Newmans Polarisationsalgorithmus, Kernighan–Lin Algorithmus sind grundsätzlich für die Analyse von statischen Netzwerken entwickelt worden. Wenn man bei dynamischen Netzwerken jeden Snapshot des Netzwerks als eigenes statisches Netzwerk behandelt und einen klassischen Algorithmus zur Gruppierung verwendet, kann es zu starken Schwankungen der Gruppenzugehörigkeiten der einzelnen Mitgliedern kommen [PART et al. 2011].

3.3.2. Qualitätsfunktionen

Um die Güte der gefundenen Communities im Netzwerk beurteilen zu können. gibt es verschiedene Qualitätsfunktionen. Diese können im Nachhinein, nachdem der Clustering Algorithmus auf den Graphen ausgeführt wurde, auf das Ergebnis angewendet werden.

3.3.2.1. Normierte Teilmenge (Normalized cut)

Die normierte Teilmenge $S \subset V$ wird definiert als

$$Ncut(S) = \frac{\sum_{i \in S,\ j \in \overline{S}} A(i,j)}{\sum_{i \in S} degree(i)} + \frac{\sum_{i \in S,\ j \in \overline{S}} A(i,j)}{\sum_{j \in \overline{S}} degree(j)}$$

Wobei *A(i, j)* die Adjazenzmatrix des Netzwerkgraphen mit den Kantengewichte zwischen den Knoten *i* und *j* der Menge *V* der Knoten des Graphen abbilden.
Somit ergibt sich die normierte Teilmenge einer Gruppe von Knoten *S*, ist die Summe der Gewichte der Kanten die mit *S* direkt aus dem Rest des Graphen verbunden sind, normalisiert zu den gesamten Kantengewichten von *S* und dem Rest des Graphen $\overline{S}$. Oder in anderen Worten, Gruppen mit kleinen *NCut(S)* sind gut vom Community Detection Algorithmus abgebildet worden, weil sie untereinander sehr stark vernetzt sind, aber nur sehr schwach mit dem Rest des Netzwerks [MEIL et al 2001].

3.3.2.2. Konduktanz und Kernighan-Lin Objectives

Die Konduktanz ist dem *Ncut(S)* sehr ähnlich. Hierbei wird wieder eine Teilmenge von Knoten $S \subset V$ aus dem Graphen analysiert.

$$Conductance(S) = \frac{\sum_{i \in S,\ j \in \overline{S}} A(i,j)}{\min\left(\sum_{i \in S} degree(i),\ \sum_{i \in \overline{S}} degree(i)\right)}$$

Der *Conductance(S)* zielt wie der *NCut(S)* auf eine starke Vernetzung in der Gruppe und eine schwache Vernetzung mit dem Rest des Netzes ab [DHIL et al 2007].

Kernighan-Lin Objectives wird hauptsächlich zur Bewertung von Algorithmen wie den Kernighan-Lin Algorithmus eingesetzt da hier Vorausgesetz wird, dass alle entstandenen Cluster dieselbe Größe aufweisen und somit versucht wird die Clusterübergreifenden Kanten zu minimieren [PART et al. 2011], [KERN et al. 1970].

$$KLObj(V_1, \dots, V_k) = \sum_{i \neq j} A\left(V_i, V_j\right) \ \mid\ |V_1| = |V_2| = \cdots = |V_k|$$

Wobei $A(V_i, V_j)$ die Summe aller Verbindungskanten zwichen den Gruppen V_i und V_j ist.

3.3.3. Community Detection Algorithmen

Die folgenden Abschnitte beschreiben einige Algorithmen zur Community Detection.

3.3.3.1. Kerninghan-Lin Algorithmus

Der Keringhan-Lin Algorithmus ist einer der klassischen Graph-Clustering Algorithmen. Er versucht dabei die Kantenteilmenge zwischen den Clustern zu minimieren und dabei die Anzahl der Knoten in den Clustern ausbalanciert zu halten. Der iterative Algorithmus beginnt immer mit einer Zweiteilung des Graphen. In jeder Iteration sucht er nach einer Untermenge an Knoten in jedem der Graphenteile, sodass ein Tauschen dieser Knoten eine Reduktion der Kantenteilmenge und somit eine größere Vernetzung in den Clustern und geringere Vernetzung zwischen den Clustern bedeutet. In anderen Worten, der Algorithmus nimmt ständig jene Knoten von der größten Partition, welche den größten Effekt bringen und teilt sie einer passenden Partition zu [KERN et al. 1970].

3.3.3.2. Zusammenfügende und teilende Algorithmen

Zusammenfügende und teilende Algorithmen operieren auf sehr ähnliche Art, arbeiten jedoch genau entgegengesetzt zueinander. Zusammenfügende Algorithmen beginnen zunächst mit einem sozialen Netzwerk, bei dem jeder Knoten als eigene Community betrachtet wird. Danach werden in jedem weiteren Schritt die Eigenschaften der Communities verglichen und Communities die passende Eigenschaften aufweisen zu einer Community zusammengeführt. Dieser Schritt wird solange wiederholt, bis keine weiteren Zusammenführungen mehr möglich sind oder die gewünschte Anzahl an Communities erreicht wurde.
Teilende Algorithmen hingegen gehen von der anderen Richtung an die Aufgabe heran. Sie gehen zunächst von einer einzigen großen Community aus, welche alle Knoten des sozialen Netzwerks beinhaltet. Danach werden in jedem Schritt die Eigenschaften der Knoten bzw. der Kanten einer Community bewertet und die Community in Subcommunities zerteilt. Dieser Vorgang wird wieder solange wiederholt, bis die gewünschte Anzahl an Communities erreicht wurde, oder keine weiteren Zerteilungen mehr möglich sind. [PART et al. 2011]

Ein Beispiel für einen teilenden Algorithmus ist der *Girvan und Newman* Algorithmus [NEWM et al. 2004] der einen *Betweenness-Wert* zwischen den Knoten als Zuordnungskriterium für die Communities verwendet. Knoten mit einem hohen *Betweenness-Wert* gehören mit einer größeren Wahrscheinlichkeit derselben Community an und bleiben somit in derselben Community. Dieser *Betweenness-Wert* kann beispielsweise der *Shortest Path* zwischen zwei Knoten sein.

Der Algorithmus arbeitet in folgenden 4 Schritten:
1. Berechne den *Betweenness-Wert* für alle Kanten im Netzwerk mit der gewählten Messmethode (z.B. *Shortest Path*)
2. Finde die Kante mit dem höchsten Wert und entferne sie vom Netzwerk
3. Berechne erneut den *Betweenness-Wert* für alle verbleibenden Kanten im Netzwerk.
4. Wiederhole ab Schritt 2 bis die gesuchte Anzahl an Communities übergeblieben ist.

3.3.3.3. Spektralclusteringalgorithmen

Spektralclusteringalgorithmen gehen von der Idee aus, einen niedrigdimensionalen Graphen mit Hilfe der k größten Eigenvektoren der Adjazenzmatrix in einen k-dimensionalen Raum einzuteilen und diesen dann mit Hilfe klassischer Clusteringtechniken, wie beispielsweise k-means, zu gruppieren [LUXB 2007]. Das Hauptproblem bei Spektralclusteringalgorithmen liegt in ihrer Komplexität. Die meisten dieser Algorithmen zur Berechnung der Eigenvektoren benötigen ein iteratives Vorgehen, bei dem zugleich in jeder Iteration komplexe Matrixmultiplikationen durchgeführt werden müssen. Durch diese sehr komplexen Berechnungen skalieren Spektralclusteringalgorithmen ab einer Anzahl von einigen zehntausend Knoten nicht mehr. Bei kleineren Graphen sind diese Algorithmen jedoch sehr genau und erzielen gute Ergebnisse. [PART et al 2011].

3.3.3.4. Multilevel Graph Partitionierung

Bei der Multilevel Graph Partitionierung wird der ursprüngliche Graph zunächst solange ausgedünnt, bis der resultierende Graph klein genug ist, eine Gruppierung der verbleibenden Knoten schnell und immer noch genau durchführen zu können. Die Schwierigkeit hierbei ist es, hierbei das richtige Maß der verbleibenden Informationen im Netzwerk zu finden. [KARY et al. 1999]
Im nächsten Schritt wird dann das Clustering mit Hilfe eines bekannten Clusteringverfahrens durchgeführt. Spektralclusteringalgorithmen eignen sich beispielsweise hierzu gut, da sie auf kleine Graphen sehr gute Ergebnisse erzielen.
Im dritten Schritt wird der Rückbau vom kleinen ausgedünnten, aber gruppierten Graphen hin zum nächsthägrößeren gemacht. Dabei werden die Knoten, die im ersten Schritt entfernt wurden, nun in die neue gruppierte Struktur eingebaut und richtig zugeordnet.
Bei sehr großen Netzwerken, wird im ersten Schritt des Algorithmus das Netzwerk chronologisch iterativ immer mehr ausgedünnt und die Zwischenergebnisse gespeichert, sodass beim Rückbau wieder Schrittweise die gespeicherten Zwischenergebnisse für die Gruppierung verwendet werden können [PART et al 2011].
Die Multilevel Graph Partitionierung ist ein schneller und sehr genauer Algorithmus, welcher sehr gute Ergebnisse im Clustering von Graphen erzielt [TENG 1999]. Somit bietet er die optimalen Bedingungen für den Einsatz zur Community Detection in soziale Netzwerke. [PART et al 2011]

3.3.3.5. Markov Clustering

Der Markov Clustering Algorithmus gruppiert einen Graphen indem er die stochastische Matrix des Graphen manipuliert [DONG 2000]. Der Algorithmus wendet dabei zwei Funktionen auf die Matrix an, *Expand* und *Inflate*. *Expand(M)* ist eine einfache Matrixmultiplikation von $M * M$. Bei der *Inflate(M, r)* Funktion wird jeder Wert in der Matrix M um einen Inflationswert $r > 1$ erhöht. Anschließen wird die Matrix so normalisiert, dass jede Spaltensumme 1 ergibt. Diese zwei Funktionen werden solange abwechselnd in Iterationen auf die Matrix angewendet, bis das gewünschte Ziel erreicht wurde.

Markov Clustering hat zwei große Nachteile. Erstens ist die *Expand* Operation eine Matrixmultiplikation, die gerade bei sehr großen Matrizen eine sehr langsame und rechenintensive Operation ist. Zweitens tendiert der Markov Clustering Algorithmus dazu sehr viele sehr kleine Cluster oder einen sehr große Cluster zu bilden was unter Umständen nicht gewünscht ist.

3.4. Knoten Klassifizierung in sozialen Netzwerken

Die vermehrte Nutzung von Online Social Networks hat zu einer Erhöhung der zur Verfügung stehenden personenbezogenen Daten im Internet geführt. Persönliche Meinungen, Gruppenmitgliedschaften, Aktivitäten, Beziehungen, und persönliche Einstellungen von Benutzern dieser Seiten sind nun im Internet verfügbar. Diese Daten sind jedoch noch sehr unstrukturiert und müssen für weitere Verarbeitungs- und Auswertungsschritte klassifiziert und typisiert werden.
Dieses Typisierungen (Labels), beinhalten politische oder religiöse Einstellungen von Usern, Hobbies, Interessensgebiete, und andere Neigungen die sich in den Profilen der Benutzer finden lassen.
Diese Labels können von verschiedenen Anwendungen genutzt werden, wie beispielsweise einer Vermittlung von Usern mit selben Interessen, Empfehlungssystem um Bücher, Videos oder andere Artikel vorzuschlagen, die den Interessen entsprechen, Expertenvermittlungssystemen bei denen bei Fragen gleich die zur Kategorie passenden Experten gefunden werden können, individualisierte Werbung für Personen mit den passenden Interesse dazu, und vieles mehr.
Benutzer können diese Label selbst setzen. Es wär wünschenswert, wenn jeder User seine Labels so wählt, dass sie perfekt zu seinen persönlichen Eigenschaften passen. Diese selbst gewählten Labels sind jedoch oft unvollständig und nicht vertrauenswürdig, bei manchen Usern fehlen sie sogar zur Gänze.
Es werden von manchen Usern absichtlich falsche Angaben gemacht, oder aus Datenschutzgründen weggelassen. Manche User vergessen auch einfach die Label ihren veränderten Interessen und Gegebenheiten anzupassen. Diese Tatsache resultiert in einer Schwächung der Effektivität der Analysemethoden. Jedoch ist das größte Problem, das Fehlen von charakteristischen Labels, die es unmöglich machen, dem User Empfehlungen zu geben [BHAG et al 2011].

3.4.1. Knoten Klassifizierungsproblem

Das oben bereits angesprochene Problem der fehlenden Labels ist auch als Knoten Klassifizierungsproblem (Node Classification Problem) bekannt, bei dessen Lösung man versucht auch jenen Knoten passende Labels zu zuweisen, die keine entsprechenden Labels vorweisen. Ein möglicher Lösungsversuch dieses Problems ist es, Experten zur Klassifizierung von Profilen einzusetzen oder entsprechende Anreizsysteme den Usern bereitzustellen, um ihre Profile passend zu klassifizieren. Diese Art der Klassifizierung wurde bereits vor Jahrzehnten von Soziologen betrieben, sie ist jedoch nicht sehr effizient und passt nicht mehr zu den heutigen Größenordnungen der sozialen online Netzwerke, bei denen es einige tausend bis millionen Nutzer geben kann.
Um diese Arbeit von Computern machen zu lassen, nutzt man heutzutage die bereits vorhandenen Informationen im partiell klassifizierten Graphen unter Zuhilfenahme von Algorithmen des maschinellen Lernens zur Klassifizierung.
Dies geschieht indem man bereits klassifizierte Knoten als Beispiele verwendet und den Klassifizierer mit diesen Beispielen trainiert, sodass er nach dem Training noch nicht klassifizierte Knoten selbstständig zuteilen kann.
Damit dies funktioniert, muss man jedoch zuerst Eigenschaften der Knoten die zu dieser Klassifizierung führen, analysieren und dem Lernprozess bereitstellen.
Es werden aber nicht nur Eigenschaften der Knoten berücksichtigt, sehr wichtig sind auch verschiedene Netzwerkeigenschaften der Knoten wie beispielsweise die Anzahl der Nachbarn des Knotens oder die erreichbaren Nachbarn nach zwei bis drei Hops, die Anzahl der Shortest Paths die durch den Knoten gehen und viele mehr.
Außerdem beeinflussen Eigenschaften von Nachbarknoten die Klassifizierung des Knoten mit. Dies hat soziologische Gründe die auf zwei soziologischen Phänomenen beruht [BHAG et al 2011]:

- Bei Personen die mit einander in Beziehung stehen und die gleichen Dinge mögen wie beispielsweise die gleiche Musik, ist es wahrscheinlicher, dass sie auch andere Interessen gemeinsam haben, zum Beispiel die selbe Literaturrichtung.
- Homophilie – Personen suchen sich andere Personen mit gleichen Eigenschaften wie sie selbst und haben somit mit diesen auch mehr Kontakt. Zum Beispiel gleiches Alter, Ausbildungsstufe, etc.

3.4.2. Methoden zur Knotenklassifizierung

Der folgende Abschnitt beschreibt zwei Arten von Methoden zur Knotenklassifizierung. Einerseits die Methoden welche lokale Klassifizierer verwenden und andererseits Random Walk basierende Methoden.

3.4.2.1. Iterative Methoden die lokale Klassifizierer verwenden

Diese iterativen Methoden machen sich die Informationen klassifizierter Knoten zu nutzen und generieren aus diesen Informationen Eigenschaften, welche sie dann dazu verwenden um lokale Klassifizierer zu erlernen. Sie bilden aus den Labels der Nachbarknoten (umgebende erreichbare Knoten) eines nichtklassifizierten Knotens, sogenannte Eigenschaftvektoren, welche sie für die weitere Berechnung als Entscheidungsbasis für die Klassifizierung verwenden. [NEVI et al. 2000]
Mit Hilfe dieser Eigenschaftsvektoren und den bekannten Labels der Knoten im Netzwerk können dann lokale Klassifizierer wie beispielsweise Decision Trees erzeugt werden. Diese Klassifizierer werden danach auf nichtklassifizierte Knoten angewendet, um ihnen passende Labels zuzuordnen [BHAG et al 2011].

3.4.2.2. Methoden die Random Walk Strategien verwenden

Diese Methoden verbreiten Labels im Graphen indem sie zufällige Wege durch den Graph gehen. Diese Methoden sind Semisupervised Machinlearning Verfahren welche versuchen, eine globale Funktion zur Knotenklassifikation über dem Graphen zu lernen. Im Gegensatz zu iterative Methoden, die Linkeigenschaften zur Berechnung verwenden um die Informationen der Nachbarn zu verwerten, benutzen Random Walk Methoden die explizite Linkstruktur um die Knoten zu klassifizieren[BHAG et al 2011], [AZRA 2007].

Die zentrale Idee der Random Walk Methoden unterliegen ist folgende:
Einem Knoten $v_i \in V$ wird der Label $c \in Y$ zugeordnet, wenn ein Random Walk beim Knoten v_i beginnt und bei einen Knoten mit dem Label c endet.Die Wahrscheinlichkeit also, dass ein Knoten $v_i \in V$ einen Label $c \in Y$ zugeordnet bekommt, ist die totale Wahrscheinlichkeit, dass ein Random Walk im Knoten v_i beginnt und an einem Knoten mit dem Label c endet. Dabei muss es für jeden nichtklassifizierten Knoten im Graphen möglich sein, mit Hilfe eines bestimmten endlichen Pfades einen anderen klassifizierten Knoten zu erreichen. Die Eintrittswahrscheinlichkeit eines Random Walk wird definiert als die Wahrscheinlichkeit p_{ij} in der *i*-ten Zeile und *j*-ten Spalte in einer Transitionmatrix *P,* die dem Pfad von Knoten v_i nach v_j zugeordnet ist. Dabei muss folgendes gelten [AZRA 2007]:

$$0 \leq p_{ij} \leq 1 \text{ und } \sum_j p_{ij} = 1.$$

3.4.2.3. Knotenklassifikation in realen sozialen Netzwerken

Eine wesentliche Herausforderung bei der Anwendung von Knotenklassifizierungsmethoden in sozialen Netzwerken ist die Größe dieser Netzwerke. Soziale Netzwerke können tausende Labels, millionen von Knoten und milliarden Kanten haben. Es wäre sehr rechenintensiv würde man Knotenklassifizierungsmethoden direkt auf diese enormen Datenmengen ansetzen [BHAG et al 2011]. Der Random Walk basierte Algorithmus von [ZHOU et al. 2009] welcher eine Matrix der Größe $n * n$ invertiert, wobei n die Anzahl an Knoten im Graphen ist, wäre für solch große Netzwerke ungeeignet. Daher gibt es verschieden Strategien, wie man dieses Problem umgehen kann. Ein Beispiel hierfür ist Random Walks zu simulieren [SARM et al. 2008] oder bei Iterativen Methoden die Zwischenergebnisse nach jeder Iteration weiterzuverwenden [MUTH et al. 1998a], [MUTH et al. 1998b].

3.5. Link Schlussfolgerungen (Link prediction)

Soziale Netzwerke sind sehr dynamisch, sie verändern sich über die Zeit hinweg, indem viele neue Knoten hinzukommen und einige wieder verschwinden. Beispielsweise Benutzer die sich neu anmelden und somit zum Netzwerk hinzugefügt werden. Diese dynamische Eigenschaft von sozialen Netzwerken macht es jedoch oft schwierig Aussagen über zukünftige Entwicklungen und Prognosen abzugeben. Eine Möglichkeit solcher Prognosen ist Link Prediction. Es wird versucht auf folgende Fragen einzugehen, um die Wahrscheinlichkeit des Auftretens zukünftiger Beziehungen zwischen zwei Knoten zu bestimmen:

- Wie wird eine Beziehung zwischen zwei Knoten von andern Knoten beeinflusst?
- Was sind die Faktoren die diesen Einfluss ermöglichen?
- Wie verändert sich die Beziehung zwischen zwei Knoten über die Zeit?
- Welche Umstände haben dazu geführt, dass eine Beziehung zwischen zwei Knoten entstanden ist?

3.5.1. Durchführung von Link Schlussfolgerungen

Gegeben ist ein soziales Netzwerk $G(V, E)$ bei dem eine Kante $e = (u, v)$ Element von E ist und eine Form der Interaktion zum Zeitpunkt $t(e)$ darstellt. Man kann mehrere Interaktionen in einen Zeitintervall aufzeichnen oder bei jeder Interaktion einen Zeitstempel verwenden, um dann den Subgraphen $G[t, t']$ generieren, der die Veränderungen beinhaltet.
Mit diesen Graphen kann man dann mit Hilfe von Supervised Learning Algorithmen zur Link Prediction ein Trainingsintervall von t_{learn} bis t_{learn}' auswählen und den Algorithmus lernen lassen. Danach kann ein zweites Testintervall ausgewählt werden, um den Lernerfolg des trainierten Modells zu testen. Dabei ist es wichtig, dass der Zeitpunkt t_{test} erst nach Ablauf des Trainingsintervalls gewählt wird.

Ist der Algorithmus gut trainiert, so kann er die auftretenden Links im Testintervall bereits im Vorhinein sehr genau berechnen und somit auch zukünftige nichtaufgezeichnete Links mit hoher Wahrscheinlichkeit vorhersagen.

Link Prediction wird oft im Internet eingesetzt; zum Beispiel für eine automatische Generierung von Hyperlinks oder im e-Commerce zur Erstellung von Kaufempfehlungssystemen. Ein weiteres Beispiel für ein Einsatzgebiet ist die Identifikation von verstecken Terroristengruppen und Kriminellen in Sicherheitsapplikationen [HASA et al. 2011].

3.6. Content Based Mining in sozialen Netzwerken

Beim Content Based Mining in sozialen Netzwerken wird nicht mehr rein auf die Struktur und den Aufbau des sozialen Netzwerks geachtet, sondern es wird auch der vom User generierte Inhalt mit in Betracht gezogen. [AGGA et al. 2011]

3.6.1. Text Mining in Sozialen Netzwerken

Soziale online Netzwerke, wie Facebook erlauben es auf einer Vielzahl verschiedener Arten Textbotschaften auszutauschen und sind normalerweise sehr reich an Text. Die User können Textkommentare auf ihre Profile posten, können Links austauschen, private Emails untereinander versenden und miteinander chatten. Aufgrund dieses reichen Angebots an Daten ist es leicht möglich den ausgetauschten Text zu analysieren und somit Information und Mehrwert aus den Daten zu schöpfen.

Obwohl es bereits eine Vielzahl unterschiedlicher Text Mining Algorithmen und Suchalgorithmen gibt, ist die Anwendung dieser in sozialen Netzwerken eine spezielle Herausforderung, da hier auch die Linkstruktur des Netzwerks Beachtung finden muss.

Die Linkstruktur in sozialen Netzwerken bietet interessante Hinweise, die zeigen, wie man dieses Problem angeht, indem man Information von einer Domain zur anderen transferieret. Diesen Algorithmus nennt man Transfair Learning.
Dabei wird Wissen in einer Domäne erhoben und über die Linkstruktur und so genannte Bridges in die andere Domäne verlagert. In sozialen Netzwerken ist das bilden dieser Bridges leicht über die vorhandenen Links möglich und somit kann der Lernprozess mittels Transfair Learning leicht eingesetzt werden. [AGGA et al. 2011]

3.6.2. Multimedia Mining in Sozialen Netzwerken

In den letzten Jahren sind digitale Videokameras und Fotoapparate immer erschwinglicher geworden, sodass diese Geräte heute bereits weit verbreitet und häufig verwendet werden. Mit diesem Trend hat sich auch die Popularität von online Foto und Videosharing Websites, wie Flickr und YouTube gesteigert in denen multimediale Informationen zwischen den Usern ausgetauscht und veröffentlicht werden können.
Mit diesen Seiten werden die multimedialen Daten auch strukturiert, wie beispielsweise in semantischen Ontologien, geordnet abgelegt. Die Untersuchung dieser strukturierten multimedialen Daten hat zu einer neuen Art der Forschung geführt, die Multimedia Informations- Netzwerks- Analyse. Multimedia Informationsnetzwerke sind sehr eng verwandt mit den Sozialen Netzwerken. Sie versuchen die Multimedialen Dateien zu verstehen, klassifizieren und den semantischen Sinn der Dateien in den Kontext einer Netzwerkstruktur darzustellen.

Ein Multimedia Informationsnetzwerk ist eine strukturierte Multimedia Sammlung in der multimediale Inhalte wie Dokumente, Fotos und Videos als Knoten dargestellt werden und durch Kanten verbunden sind. Sie können somit als eine Kombination von multimedialen Inhalten mit einem sozialen Netzwerk gesehen werden. Diese Kanten können wie bei anderen Netzwerken als logische Beziehungen zwischen den Knoten gesehen werden. Beispiele solcher Kanten sind Meta Informationen über den User, Ortsbezogene Informationen oder Tags. Aber auch herkömmliche Soziale Netzwerk Seiten beinhalten Multimediale Daten; so hostet Facebook milliarden von Bilddateien seiner Usern, welche untereinander durch Hyperlinks, Beziehungen von Usern, oder Tags in Beziehung stehen.

Eine wichtige Herausforderung bei der Analyse von Multimedia Daten in sozialen Netzwerken ist das Zusammenspiel zwischen der Netzwerkstruktur und den multimedialen Informationen in diesen Netzwerken. Diese Kombination von Struktur und Information bildet einen weitaus reicheren Informationsgehalt, als sie jede Komponente für sich alleine nie bieten könnte. Die Netzwerkstruktur und das darauf basierende semantische Wissen kann hierbei genutzt werden, um diese Inhalte, in für Computer verständliches Wissen umzuwandeln.

Somit ist es wichtig, nicht nur die multimedialen Inhalte für sich zu analysieren, sondern sie im Kontext des sozialen Netzes zu erforschen, da sich nur so die Informationen beider Komponenten ergänzen. [LIAN et al. 2011]

3.7. Visualisierung von sozialen Netzwerken

Die Visualisierung von sozialen Netzwerken wurde in den letzten Jahren immer mehr zu einem wichtigen Werkzeug um die inneren Vorgänge und die dynamischen Strukturen von sozialen Netzwerken zu analysieren. Seit dem Zeitpunkt, an dem die ersten Soziogramme verwendet wurden, haben sich die sozialen Netzwerke und vor allem die damit verbundenen Datenmengen, extrem verändert. Nach heutigen Stand der Technik würde das bloße betrachten von statistischen Listen als unpassend angesehen werden. Es existiert eine Vielzahl an verschiedenen Visualisierungsmethoden für soziale Netzwerke. Die meisten jedoch sind Variationen der ursprünglichen Soziogrammart, bei der Akteure in einem Netzwerk mit graphischen Elementen und Beziehungen zwischen diesen Akteuren mittels Verbindungslinien zwischen diesen Elementen dargestellt werden.
Diese Art der Darstellung hat sich im Laufe der Zeit als leicht verständlich und einfach realisierbar herausgestellt und bietet detaillierte Informationen über aktuelle Beziehungen von Akteuren und deren Position im Graphen.
Frühe Visualisierungstools wurden hauptsächlich für Forschungszwecke entwickelt und eingesetzt, mit dem Aufschwung von Social Networking Sites wurden jedoch neue Tools zur kommerziellen Anwendung entwickelt.
Des Weiteren gibt es eine Menge an Tools die Visualisierungs- und Analyseverfahren kombiniert anbieten, wie beispielsweise *JUNG* oder *Pajek*. (siehe [CARL et al. 2011])
Viele Visualisierungstools sehen die Analyse das Clustern und das Filtern als einen Vorprozess, der vor der Visualisierung gemacht wird, um die Visualisierung durchführen zu können. [CARL et al. 2011]

3.7.1. Arten der Visualisierung

Es gibt vier Arten von Taxonomien der Visualisierung die sehr häufig verwendet werden, die strukturelle Visualisierung, die semantische und zeitliche Visualisierung und die statistische Visualisierung [CARL et al. 2011].
Die strukturelle Darstellung wird meist dazu verwendet die Beziehungen einzelner User darzustellen. Diese Darstellungsform bietet einen genauen Einblick von einem bestimmten Ausschnitt des Netzwerks; beispielsweise die Detailansicht einer bestimmten Gruppe. Sie ist jedoch zur Darstellung eines gesamten Netzwerks von mehreren tausend bis millionen Knoten ungeeignet.
Die semantische Darstellungsform wird oft dazu verwendet den User einen Überblick über ein Netzwerk bzw. bestimmte Aspekte des Netzwerks zu ermöglichen. So kann beispielsweise die Gruppenstruktur eines Sozialen Netzwerks semantisch dargestellt werden, ohne dabei auf die genauen, viel zu feinen Strukturen des Netzwerks eingehen zu müssen.
Die zeitliche Visualisierung wird dazu verwendet zeitliche Abläufe im sozialen Netzwerk sichtbar zu machen.
Es gibt verschiedene Arten der statistischen Visualisierungen die beispielsweise Entwicklungen bestimmter Netzwerkkennzahlen grafisch veranschaulichen. [CARL et al. 2011]

4. Zusammenfassung und Ausblick

In dieser Arbeit wurden zunächst die Grundbausteine sozialer Netzwerke beschrieben. Danach wurden statische und dynamische Methoden zur strukturellen Analyse von sozialen Netzwerken erklärt, welche es ermöglichen, Kennzahlen über das Netzwerk zu berechnen. Da sich soziale Netzwerke über die Zeit hinweg verändern, werden mittels dynamischer Methoden, Werte für das dynamische Verhalten der Netzwerke berechnet, während statische Methoden lediglich auf Snapshots, also Momentaufnahmen der sozialen Netzwerke angewandt werden.
Des Weiteren wurde gezeigt was Community Detection ist, wie es eingesetzt werden kann und mit welchen Methoden Communities im Netzwerk gefunden werden können.
Danach wurde beschrieben, warum es Knotenklassifizierung gibt und wie das Knoten Klassifizierungsproblem gelöst werden kann.
Bei der Link Schlussfolgerung wird versucht, mit Hilfe von vorhandenem Wissen aus dem Netzwerk, zukünftige Links zwischen Knoten vorherzusagen und dies wirtschaftlich, zum Beispiel für Werbung zu nutzen. Zu guter Letzt wurde das Thema Visualisierung behandelt. Visualisierung hat, in den letzten Jahren in der sozialen Netzwerk Analyse durch die steigende Anzahl an Knoten im Netzwerk, immer mehr an Bedeutung gewonnen und ist momentan ein Gebiet an dem viel geforscht und entwickelt wird.
Gerade zur heutigen Zeit wird das Data Mining in sozialen Netzwerken immer wichtiger werden. Es stehen immer mehr Daten der einzelnen Personen zur Verfügung und große Unternehmen, wie Facebook und Co., wollen diese Daten bestmöglich vermarkten. In Bereichen wie Multimedia Mining, welches noch relativ jung ist, kann noch viel geforscht und verbessert werden. Des Weiteren ist ein Trend bei den Usern in sozialen Netzwerken erkennbar, die immer mehr multimediale Inhalte produzieren und in sozialen Netzwerken veröffentlichen. Durch das Analysieren der Daten im sozialen Netzwerk, ist es schon heute sehr gut möglich, personalisierte Werbung zu generieren und diese maßgeschneiderte Werbung wird in Zukunft noch viel verstärkter zum Einsatz kommen.

5. Literaturverzeichnis

[AGGA 2011] Charu C. Aggarwal - Sozial Network Data Analytics; Springer Verlag 2011

[AGGA et al. 2011] Charu C. Aggarwal, Haixun Wang - Text Mining in Social Networks; Social Network Data Analytics; Springer Verlag 2011

[AZRA 2007] A. Azran. - The rendezvous algorithm: Multiclass semi-supervised learning with markov random walks; ICML 2007.

[BARA et al. 2003] A.L. Barabási and E. Bonabeau - Scale-free networks; Scientifc American 288(5):60 2003.

[BHAG et al 2011] Smrit Bhagat, Graham Cormode, S. Muthukrishan Node Classification in Social Networks; Springer Verlag 2011

[CARL et al. 2011] Carlos D. Correa, Kwan-Liu Ma - Visualizing social Networks, Social Network Data Analystics; Springer Verlag 2011

[DHIL et al 2007] I.S. Dhillon, Y. Guan, and B. Kulis - Weighted Graph Cuts without Eigenvectors: A Multilevel Approach. IEEE Trans. Pattern Anal. Mach. Intell., 29(11):1944–1957, 2007.

[DOMI et al. 2001] P. Domingos, M. Richardson - Mining the network value of customers. In KDD ’01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pages 57–66. ACM, 2001

[DONG 2000] S. Van Dongen. - Graph Clustering by Flow Simulation. PhD thesis, University of Utrecht, 2000

[ELLI et al. 2007] Ellison NB, Steinfield C, Lampe C - The benefits of Facebook ‘‘friends:’’ social capital and college students’ use of online social network sites; J Comput Mediat Commun 12(4): 1143–1168 2007

[HASA et al. 2011] Mohammad Al Hasan, Mohammed J. Zaki - A survey of link prediction in social Networks Springer Verlag 2011

[KARY et al. 1999] G. Karypis and V. Kumar - A fast and high quality multilevel scheme for partitioning irregular graphs; SIAM Journal on Scientific Computing, 20, 1999.

[KEMP et al 2003] D. Kempe, J. Kleinberg, and É. Tardos - Maximizing the spread of influence through a social network. In KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 137–146, New York, NY, USA, ACM 2003

[KERN et al. 1970] B. Kernighan and S. Lin - An Effcient Heuristic Procedure for partitioning graphs. The Bell System Technical J., 49, 1970

[LACK et al. 2009] Lackaff D, Lim D, Kwon KH, Tripoli A, Stefanone MA - Resource mobilization on social network sites. Paper presented at the annual conference of the National Communication Association, Hilton Chicago, Chicago, IL, USA 2009

[LESK et al. 2007] J. Leskovec, L.A. Adamic, and B.A. Huberman - The dynamics of viral marketing; ACM Transactions on the Web (TWEB), 1(1):5, 2007

[LESK et al. 2005] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos - Graphs over time: densifcation laws, shrinking diameters and possible explanations. In Proc. of ACM SIGKDD, pages 177–187, Chicago, Illinois, USA, ACM Press. 2005

[LIAN et al. 2011] Liangliang Cao, GuoJun Qi, Shen-Fu Tsai, Min-Hsuan Tsai, Andrey Del Pozo, Thomas S. Huang, Xuemei Zhang und Suk Hwan Lim - Multimedia Information Networks in social media; Springer Verlag 2011

[LUXB 2007] U. Von Luxburg - A tutorial on spectral clustering. Statistics and Computing, 17(4):395–416, 2007.

[MCGL et al. 2011] Mary McGlohon Laman Akoglu Christos Faloutsos - Statistical properties of social networks Springer Verlag 2010

[MEIL et al 2001] M. Meila and J. Shi - A random walks view of spectral segmentation. AI and Statistics (AISTATS), 2001

[MUTH et al. 1989a] S. Muthukrishnan, B. Ghosh, and M. H. Schultz - First- and second-order diffusive methods for rapid, coarse, distributed load balancing. Theory Comput. Syst., 31(4), 1998

[MUTH et al. 1998b] S. Muthukrishnan and T. Suel. - Second-order methods for distributed approximate single- and multicommodity Row. In RANDOM, 1998

[NEVI et al. 2000] J. Neville and D. Jensen. Iterative classi¿cation in relational data. In Workshop on Learning Statistical Models from Relational Data, AAAI, 2000.

[NEWM et al. 2004] M.E.J. Newman and M. Girvan - Finding and evaluating community structure in networks. Phys. Rev. E, 69(2):026113, Feb 2004.

[PART et al. 2011] S. Parthasarathy, Y. Ruan, V. Satuluri – Community discovery in social Networks. Applications, Methods and Emerging Trends The Ohio State University ; Springer 2011

[ROSE 2010] D. Rosen, G. Barnett J. Hyun Kim - Social networks and online environments: when science and practice co-evolve; Springer Verlag 2010

[STEF et al. 2010] Stefanone MA, Lackaff D, Rosen D - The relationship between traditional mass media and 'social media': reality television as a model for social network site behavior. J Broadcast Electron Media, 54(3):508–525 2010

[SCOT 2010] John Scott - Social network analysis: development, advances, and prospects; Springer Verlag 2010

[SARM et al. 2008] A. D. Sarma, S. Gollapudi, and R. Panigrahy - Estimating pagerank on graph streams. In PODS, 2008.

[TENG 1999] S.H. Teng - Coarsening, sampling, and smoothing: Elements of the multilevel method. Algorithms for Parallel Processing, 105:247–276, 1999.

[WASS et al. 1994] Wasserman S, Faust K 1994 - Social structure and opinion formation. HP Labs Research Paper. Palo Alto.
http://www.hpl.hp.com/research/idl/papers/opinions/opinions.pdf . 4 April 2010

[ZHOU et al. 2009] Y. Zhou, H. Cheng, and J. X. Yu - Graph clustering based on structural/attribute similarities. In VLDB, 2009.