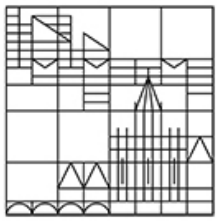

Fachbereich Informatik und Informationswissenschaften
Lehrstuhl: Prof. Dr. Ulrik Brandes
Seminar: Algorithmische Spieltheorie
Betreuer: Bobo Nick
Sommersemester 2009

Diffusion im sozialen Netzwerk

Christine Mellau

Basierend auf:
N. Nisam, T. Roughgarden, E. Tardos, V. Vazitani:
Algorithmic Game Theory

Kapitel 24
Jon Kleinberg:
Cascading Behavior in Networks: Algorithmic and Economic Issues

christine.mellau(at)googlemail.com

Inhaltsverzeichnis

1 Einleitung

Diese Ausarbeitung behandelt die Ausbreitung von Informationen und Verhalten in sozialen Netzwerken. Diese Diffusionsprozesse können unterschiedlichster Art sein, wie zum Beispiel: Ausbreitung einer Religion, politische und soziale Bewegungen wie Frauenbewegung, Sklavenbefreiungsbewegung, Anti-Atomkraftbewegung, die Verbreitung eines Instant Messengers, die Durchsetzung von Innovationen (wie Handys), plötzlicher Erfolg eines Produkts (z.B. iPod) oder die Entwicklung eines Buches zum Bestseller. Ein aktuelles Beispiel wäre auch die Ausbreitung des Verhaltens sich gegen die Schweinegrippe impfen zu lassen. Die Ausbreitung des Impfstoffes ist auch ein sozialer Prozess, weil sich die Menschen an den Erfahrungen ihres Umfeldes orientieren.

Man ist sich schon seit längerem dieser Prozesse grundsätzlich bewusst, aber ihre Erforschung begann erst Mitte des 20. Jahrhunderts innerhalb der Soziologie. Die ersten Untersuchungen waren empirisch (z.B Coleman et al. 1966, Rogers 1995). Seit den 70er Jahren beteiligen sich auch Ökonomen, Mathematiker und Informatiker an der Erforschung der Prozesse indem sie Modelle dafür entwickeln.

Zuerst stelle ich einige Diffusionsmodelle vor, anschließend betrachten wir eine aktuelle empirische Studie um dann mit einem Ausblick auf offene Forschungsaufgaben zu schließen.

2 Ein erstes spieltheoretisches Modell

Für dieses erste Modell machen wir folgende beiden Annahmen: Zwischen zwei in Beziehung stehenden Individuen existiert der Anreiz in ihrem Verhalten übereinzustimmen. Je mehr sich ein neues Verhalten im Umfeld eines Individuums bereits durchgesetzt hat, desto wahrscheinlicher wird dieses Individuum das neue Verhalten ebenfalls annehmen. Unter den in Abschnitt 1 genannten Diffusionsprozessen würde zum Beispiel die Ausbreitung eines Instant-Messengers wie ICQ diesen Annahmen gerecht.

2.1 Das Modell

Das soziale Netzwerk wird hier wie auch im Folgenden durch einen Graphen $G = (V, E)$ modelliert. Die Knoten V repräsentieren die Individuen und zwischen den Individuen die auf irgend eine Art und Weise sozial miteinander verbunden sind, befinden sich Kanten E. Jeder Knoten ist entweder in Zustand A (altes Verhalten) oder in Zustand B (neues Verhalten). Der Anreiz an einer Kante (v, w) im Verhalten übereinzustimmen wird durch ein Spiel zwischen v und w mit der Auszahlungsmatrix

	A	B
A	q	0
B	0	$1-q$

modelliert. Jeder Knoten spielt dieses Spiel mit jedem seiner Nachbarn. Der Gesamtgewinn eines Knotens ist die Summe seiner Einzelgewinne. Schauen wir uns an, welche Strategie sich für dieses Spiel ergibt und wie der Parameter $q \in (0,1)$ die Diffusion beeinflusst. Sei d_v die Anzahl der Nachbarn von Knoten v wobei d_v^A in Zustand A und d_v^B in Zustand B sind. Wählt v Zustand A, beträgt der Gewinn $q \cdot d_v^A$ und wählt V Zustand B dann $(1-q) \cdot d_v^B$. Also ist B die bessere Wahl, wenn $d_v^B > qd_v$ und A wenn $qd_v > d_v^B$. Der Eindeutigkeit halber möge v bei $d_v^B = qd_v$ Zustand B wählen. Der Parameter q wirkt sich als Schwelle aus, der angibt welcher Anteil der Nachbarn B angenommen haben muss, damit der Wechsel auf B vorteilhaft ist. Demnach breitet sich ein Verhalten mit kleinem q leichter aus als ein Verhalten mit großem q.
Dieser Strategie folgend aktualisieren die Knoten in den diskreten Zeiten $t = 1, 2, 3, \ldots$ ihren Zustand abhängig von den Zuständen ihrer Nachbarn. Sei S die Menge der Knoten die sich anfänglich in Zustand B befinden und $h_q^k(S)$ die Menge der Knoten in Zustand B nach k Runden mit Schwelle q. Im Beispiel der Ausbreitung von ICQ kann jeder das Programm (neu-)installieren aber auch später wieder deinstallieren. Solche Diffusionsvorgänge in denen die Knoten von Verhalten A zu B aber auch zurück von B nach A wechseln können nennt man regressiv. Im regressiven Fall ist zu beachten dass im Allgemeinen nicht $S \subset h_q^1(S)$ gilt. Oft kommen aber auch progressive Diffusionsvorgänge vor in denen nur ein Wechsel von A nach B möglich ist. Wir nennen einen Knoten aktiv wenn er in Zustand B ist.

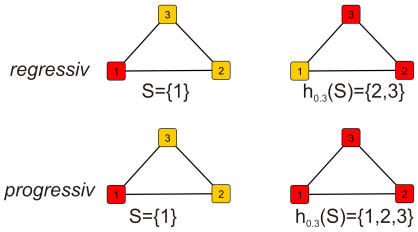

Abbildung 1: Ein Schritt eines regressiven und progressiven Diffusionsvorgangs

Wir wissen bereits, dass sich ein Verhalten schwerer ausbreitet wenn die Schwelle q groß ist. Daher stellt sich die Frage wie groß die Schwelle q sein kann, sodass immer noch eine Infektionsmenge d.h eine endliche Startmenge S existiert, von der ausgehend irgendwann alle Knoten das neue Verhalten B angenommen haben. Die maximale derartige Schwelle heißt Infektionsschwelle. Wobei wir annehmen, dass die Knotenmenge V abzählbar unendlich ist (sonst existiert immer die triviale Infektionsmenge $S = V$) und jeder Knoten endlich viele Nachbarn hat. Die Höhe der Infektionsschwelle hängt allein von der Topologie des Graphen ab. Je höher die Infektionsschwelle ist, desto mehr Potential steckt in dem Graphen, dass eine Startmenge S letztendlich den ganzen Graphen erreicht.

Beispiel 1: Der Graph G ist ein in beide Richtungen unendlicher Pfad. Wir betrachten einen regressiven Prozess mit Infektionsschwelle $q = 1/2$. Abbildung 2 zeigt die Ausbreitung eines neuen Verhaltens für die Startmenge $S = \{0\}$. Im ersten Schritt nehmen die Nachbarn von 0 das neue Verhalten an, während Knoten 0 keinen Nachbarn mit

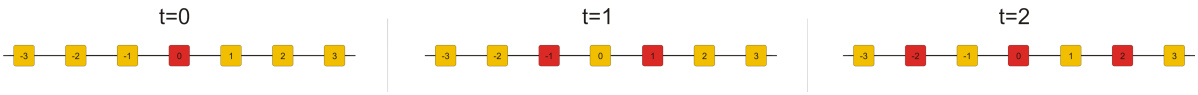

Abbildung 2: Zwei Schritte eines regressiven Prozesses mit $S = \{0\}$ und $q = 1/2$

Verhalten B hat und auf A zurückfällt. Im Folgenden werden zur Zeit t die Knoten t und $-t$ aktiv, gemeinsam mit den dazwischenliegenden geraden bzw. ungeraden Knoten. Das System pendelt zwischen geraden und ungeraden Knoten aber kein Knoten bleibt dauerhaft aktiv.

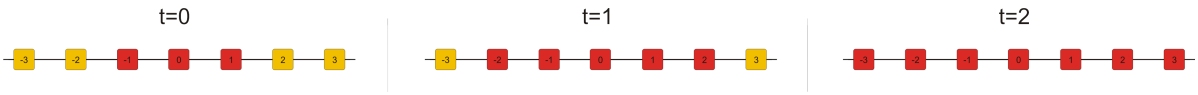

Abbildung 3: Zwei Schritte eines regressiven Prozesses mit $S = \{-1, 0, 1\}$

In Abbildung 3 verändern wir die Startmenge auf $\tilde{S} = \{-1, 0, 1\}$. Knoten 2 und -2 werden im ersten Schritt aktiv und kein Knoten fällt zurück.Allgemein sind $\{-(t+1), -t, \ldots, t, t+1\}$ zur Zeit t aktiv. $\tilde{S} = \{-1, 0, 1\}$ ist im Gegensatz zu $S = \{0\}$ infektiös. Also ist die Infektionsschwelle von G mindestens $1/2$. Tatsächlich ist die Infektionsschwelle genau $1/2$, denn bei jedem $q > 1/2$ und jeder endlichen Menge S kann sich B nicht weiter als der äußerste Knoten von S ausbreiten. Versucht man einen Graphen mit einer Infektionsschwelle $q > 1/2$ zu finden, so mag das nicht gelingen. Es stellt sich die Frage: Gibt es einen Graphen G mit Infektionsschwelle $q > 1/2$?
Dies wollen wir im Folgenden beantworten.

Satz 2.1 (Morris 2000)**.** *Für jeden Graphen G existiert eine endliche Infektionsmenge S zur Schwelle q bezüglich dem regressiven Prozess genau dann wenn eine solche bezüglich dem progressiven Prozess existiert.*

Mit anderen Worten: Progressives und regressives Modell haben die selbe Infektionsschwelle. Der Satz mag überraschend kommen, denn intuitiv scheint sich ein progressiver Prozess leichter auszubreiten als ein regressiver, schließlich können einmal aktive Knoten ihren Zustand nicht mehr verlieren.

Beweis. Bezeichne h_q^k die Menge der aktiven Knoten nach k Runden mit Schwelle q bezüglich des regressiven Prozesses und $\bar{h}_q^k$ bezüglich des progressiven Prozesses. Wenn eine Infektionsmenge bezüglich des regressiven Prozesses existiert, dann ist diese offensichtlich auch Infektionsmenge bezüglich des progressiven Prozesses. Zu zeigen bleibt die Umkehrung, dass wenn eine Infektionsmenge für den progressiven Prozess existiert, auch eine für den regressiven Prozess existiert.
Sei S Infektionsmenge bezüglich des progressiven Prozesses $\bar{h}_q^k$. Da alle Knoten in G endlich viele Nachbarn haben, ist S zusammen mit allen Knoten die einen Nachbarn in S haben die endliche Menge $\tilde{S}$. Da S Infektionsmenge, wird $\bar{h}_q^k$ irgendwann alle Knoten von G beinhalten. Insbesondere gibt es ein l sodass $\bar{h}_q^l \supset \tilde{S}$.

Wir zeigen, dass $T := \bar{h}_q^l(S)$ Infektionsmenge bezüglich des regressiven Prozesses ist.
Behauptung: $\forall X \subset V \; \forall j \geq 1 : \bar{h}_q^j(X) = X \cup h_q^1(\bar{h}_q^{j-1}(X))$
Diese zeigen wir mittels Induktion nach j.
IA: $j = 1$: $\bar{h}_q(X) = X \cup h_q(X)$✓
Ein progressiver Schritt unterscheidet sich vom regressiven nur dadurch, dass die aktive Knotenmenge X beibehalten wird.
IS:

$$\begin{aligned}\bar{h}_q^{j+1}(X) = \bar{h}_q(\bar{h}_q^j(X)) &\overset{IA}{=} \bar{h}_q^j(X) \cup h_q(\bar{h}_q^j(X)) \\ &\overset{IV}{=} X \cup h_q(\bar{h}_q^{j-1}(X)) \cup h_q(\bar{h}_q^j(X)) = X \cup h_q(\bar{h}_q^j(X))\end{aligned}$$

Sei $j > l$ dann gilt mit der Behauptung: $\bar{h}^j(S) = S \cup h(\bar{h}^{j-1}(S))$
T ist in $\bar{h}^{j-1}(S)$ enthalten und folglich auch alle Nachbarn von S. Das bedeutet S ist von aktiven Nachbarn eingeschlossen und kann in einem regressiven Schritt nicht herausfallen. Es gilt $S \subset h(\bar{h}^{j-1}(S))$ und wir können $\bar{h}^j(S) = h(\bar{h}^{j-1}(S))$ schreiben.
Mit Induktion folgt für $j > l$: $h^{j-l}(T) = h^{j-l}(\bar{h}^l(S)) = \bar{h}^j(S)$.
Folglich ist T Infektionsmenge des regressiven Prozesses. □

Wir zeigen an einem Beispiel das Beweisprinzip. Betrachten wir erneut einen unendlichen Pfad mit $q = 1/2$:

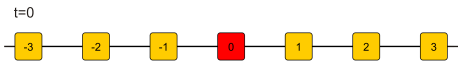

Abbildung 4: Startzustand

Aus Beispiel 1 wissen wir, dass $S = \{0\}$ nicht infektiös ist bezüglich des regressiven Prozesses. Allerdings ist S infektiös bezüglich des progressiven Prozesses. Um aus S eine Infektionsmenge T bezüglich des regressiven Prozesses zu machen, lassen wir den progressiven Prozess solange laufen, bis S vollständig eingehüllt wird. Das ist schon nach einer Runde der Fall.

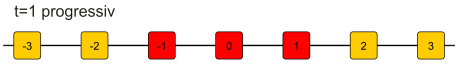

Abbildung 5: Eingehüllte Startmenge

Diese größere Menge $T = \bar{h}_{0.5}^1(S) = \{-1, 0, 1\}$ ist, wie im Beweis gezeigt wird, robust genug, dass die Ausbreitung mit dieser Startmenge auch regressiv den ganzen Graphen erreicht.

Abbildung 6: Regressiver Prozess mit Startmenge $T = \{-1, 0, 1\}$

Progressiver und regressiver Prozess verlaufen von hier an gleich, es gilt $h_{0.5}^{j-1}(T) = h_{0.5}^{j-1}(\bar{h}_{0.5}(S)) = \bar{h}_{0.5}^{j}(S)$.

Mit Satz 2.1 sind wir der Beantwortung der Frage ob es einen Graphen mit Infektionsschwelle $q > 1/2$ gibt ein Stück näher gekommen. Weil die Antwort die selbe ist, egal ob wir einen regressiven oder progressiven Prozess betrachten.

Satz 2.2 (Morris 2000)**.** *Die Infektionsschwelle eines Graphen mit abzählbar unendlicher Knotenmenge und endlichen Knotengraden ist höchstens* 0.5 .

Beweis. Sei $q > 1/2$ und S endliche Teilmenge von V. Wir zeigen: S ist nicht Infektionsmenge des progressiven Prozesses.
Für eine leichtere Notation schreiben wir S_j für $\bar{h}_q^j(S)$. Sei $\delta(X)$ die Menge der Kanten mit einem Ende in X und dem anderen Ende nicht in X. Sei $d(X) := |\delta(X)|$
Behauptung: $\forall j > 0$ mit $S_{j-1} \subsetneq S_j$ ist $d(S_j) < d(S_{j-1})$
Sei $v \in S_j - S_{j-1}$ d.h v wurde in Runde j aktiv.
Von v nach S_{j-1} gehende Kanten zählen zu $d(S_{j-1})$ aber nicht zu $d(S_j)$. Von v nach $S_j - S_{j-1}$ gehende Kanten zählen weder zu $d(S_{j-1})$ noch zu $d(S_j)$. Von v nach $V - S_j$ gehende Kanten zählen zu $d(S_j)$ aber nicht zu $d(S_{j-1})$. Da $q > 1/2$ hatte v zur Zeit $j-1$ mehr aktive Nachbarn als inaktive d.h mehr Kanten nach S_{j-1} als nach $V - S_{j-1}$. Insbesondere hatte v mehr Kanten nach S_{j-1} als nach $V - S_j$. Summieren wir die Beiträge von allen Kanten von allen Knoten aus $S_j - S_{j-1}$ zu $d(S_j)$ und $d(S_{j-1})$ erhalten wir $d(S_j) < d(S_{j-1})$. Aus der Behauptung folgt dass die Zahlenfolge $(d(S_j))_j$ streng monoton fallend ist solange die Mengenfolge $(S_j)_j$ streng monoton wachsend ist. Aber da nach Voraussetzung alle Knoten endlichen Knotengrad haben und S_j für alle j endlich ist, ist $d(S_j)$ für alle j eine natürliche Zahl und somit nach unten beschränkt. Daraus folgt dass ein $k \in \mathbb{N}$ existiert mit $S_k = S_{k+1} = S_{k+2} = \dots$. Da S_k endlich ist kann S_k nicht gleich V sein und daher S nicht Infektionsmenge.

□

3 Weitergehende Modelle

Für viele Diffusionsvorgänge in sozialen Netzwerken liefert das bisherige Modell ein ungenügend realistisches Abbild. Dafür haben wir zu viele, stark vereinfachende Modellannahmen gemacht. Wir wollen eine größere Heterogenität im sozialen Netzwerk erfassen. Bisher hatte jedes Individuum die selbe Schwelle q, aber im Allgemeinen haben Individuen unterschiedliche Empfänglichkeiten für das neue Verhalten. Auch die gleiche Gewichtung der Nachbarn soll aufgehoben werden. Weiter gab es in unserem Einstiegsmodell nur symmetrische Beziehungen. Das neue Modell soll berücksichtigen, dass der Einfluss von v auf w ein anderer sein kann als der von w auf v.
Von nun an betrachten wir stets einen progressiven Prozess auf einem endlichen oder unendlichen Graphen G.

3.1 Lineares Schwellenmodell

Das soziale Netzwerk sei ein gerichteter Graph. Jede Kante (v, w) habe eine Gewichtung $b_{vw} \in [0, 1]$, welche die Stärke des Einflusses von v auf w repräsentiert. Im gerichteten Graphen ist w ein Nachbar von v wenn es eine Kante (w, v) von w nach v gibt. Sei $N(v)$ die Menge der Nachbarn von v. Wir fordern $\sum_{w \in N(v)} b_{wv} \leq 1$. Wie zuvor hat jeder Knoten eine Schwelle, die angibt welcher Teil der Nachbarn das neue Verhalten angenommen haben muss bevor ein Knoten Verhalten B annimmt. Allerdings hat jeder Knoten v eine eigene Schwelle $\theta_v := \theta(v)$ wobei θ gleichverteilte Zufallsvariable auf $[0, 1]$ ist. In diskreten Zeiten $t = 1, 2, \ldots$ wird ein Knoten v aktiv wenn die gewichtete Summe seiner aktiven Nachbarn mindestens θ_v erreicht: $\sum_{\text{aktives } w \in N(v)} b_{wv} \geq \theta_v$

Beispiel 2:

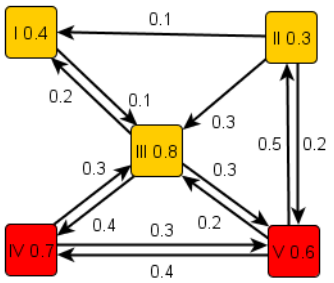

Abbildung 7: Startzustand $t = 0$

Der Graph aus Abbildung 7 umfasst 5 Knoten von denen nur IV und V zu Beginn Verhalten B haben. In jedem Knoten steht dessen Schwelle θ und an den gerichteten Kanten die Gewichtungen. Alle Gewichtungen sind zulässig, denn für jeden Knoten v gilt $\sum_{w \in N(v)} b_{wv} \leq 1$. Mit einer Tabelle ermitteln wir welche Knoten in $t = 1$ aktiv werden.

Knoten	$\sum$	θ	aktiv?
I	0	0.4	nein
II	0.5	0.3	ja
III	$0.3 + 0.2$	0.8	nein

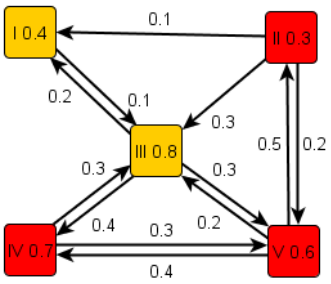

Abbildung 8: $t = 1$

Betrachten wir t=2

Knoten	$\sum$	θ	aktiv?
I	0.1	0.4	nein
III	$0.3 + 0.2 + 0.3$	0.8	ja

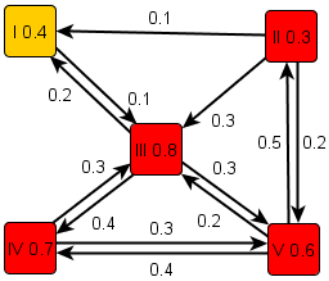

Abbildung 9: $t = 2$

Damit ist der Prozess zu Ende, da Knoten I nicht aktiviert werden kann.

3.2 Allgemeines Schwellenmodell

Im linearen Schwellenmodell waren die Beeinflussungen strikt additiv. Im allgemeinen Schwellenmodell heben wir diese Einschränkung auf. Es soll auch möglich werden Regeln folgender Art zu formulieren: Ein Individuum nimmt Verhalten B an, wenn ein Verwandter und ein Bekannter dies getan haben.
Wir behalten vom linearen Schwellenmodell bei, dass G ein gerichteter Graph und jeder Knoten eine eigene Schwelle θ_v besitzt, wobei θ gleichverteilt auf $[0, 1]$ ist. Ein Knoten wird aktiv, wenn die Beeinflussung der aktiven Nachbarn diese Schwelle erreicht. Neu ist, dass die Beeinflussung eines Knotens v durch eine Funktion $g_v(X) : N(v) \to [0, 1]$ gegeben ist. Knoten v wird aktiv, wenn die Menge X seiner aktiven Nachbarn $g_v(X) \geq \theta_v$ erfüllt. Das allgemeine Schwellenmodell erfasst jede Ausbreitung mit Schwellencharakter. Dabei ist die Annahme, dass θ gleichverteilt ist, keine Einschränkung, denn andere Verteilungen können durch modifizieren der Funktionen g_v repräsentiert werden.

Beispiel:

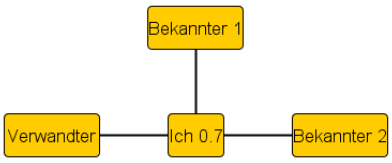

Abbildung 10: Allg. Schwellenmodell

$$g(X) = \begin{cases} 0.5, X = \{V\} \\ 0.1, X = \{B1\} \\ 0.2, X = \{B2\} \\ 0.8, X = \{V, B1\} \\ 0.8, X = \{V, B2\} \\ 0.4, X = \{B1, B2\} \\ 0.9, X = \{V, B1, B2\} \end{cases}$$

Die Funktion g gibt den Einfluss der aktiven Knoten auf mein Verhalten an. Weder meine Bekannten noch mein Verwandter könnten mich alleine aktivieren. Ich würde erst das neue Verhalten annehmen wenn mein Verwandter und mindestens ein Bekannter dies getan haben, denn $g(\{V, B1\}) = (\{V, B2\}) = 0.8 \geq 0.7$ und $g(\{V, B1, B2\}) = 0.9 \geq 0.7$ sind größergleich als meine Schwelle. Die Beeinflussung in all diesen Fällen ist nicht additiv z.B. $g(\{V\}) + g(\{B1\}) \neq g(\{V, B1\}$.

3.3 Kaskadenmodell

Bisher beruhten die Modelle auf einer schwellenförmigen Ausbreitung. Indem einige das neue Verhalten annahmen wurden Schwellen anderer erreicht, welche es dann auch annahmen und so weiter. Das Kaskadenmodell basiert dagegen auf der Vorstellung, dass das neue Verhalten ansteckend ist und Individuen es sich wie bei einer Epidemie „einfangen“. Im Kaskadenmodell haben wir einen gerichteten Graphen dessen Kanten Übertragungswahrscheinlichkeiten p_{uv} besitzen. Sei Knoten v aktiv und sei (v, u) eine Kante mit u nicht aktiv, so hat v genau einen Versuch mit Erfolgswahrscheinlichkeit p_{uv} um u mit dem neuen Verhalten anzustecken. Wenn dies gelingt hat u die Chance seine Nachbarn anzustecken und so breitet sich das Verhalten aus.

Beispiel 3:

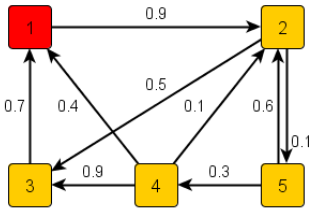

Abbildung 11: Kaskadenmodell

Zu Beginn ist nur Knoten 1 aktiv. Knoten 1 hat einen Bernoulliversuch mit Erfolgswahrscheinlichkeit 0.9 um Knoten 2 zu infizieren. Gelingt dies hat Knoten 2 die Chancen Knoten 3 mit Erfolgswahrscheinlichkeit 0.5 und Knoten 5 mit Erfolgswahrscheinlichkeit 0.1 zu infizieren.

4 Die Suche nach einflussreichen Knoten

Angenommen wir sind eine Firma die ein Produkt durch Mundpropaganda verbreiten möchte. Unsere Strategie wäre, Daten über das soziale Netzwerk unserer potenziellen Kunden zu sammeln und dann eine Menge S anfänglicher Kunden auszuwählen, an die wir das Produkt direkt vermarkten. Um uns dann drauf zu verlassen, dass diese das Produkt annehmen und durch ihren Einfluss die Ausbreitung des Produkts auslösen.
Angenommen wir wählen zur Beschreibung des Diffusionsprozesses das Kaskadenmodell. Dann müssten wir aus den gesammelten Daten ein Kaskadenmodell entwerfen, was an sich schon eine Herausforderung ist. An dieser Stelle wollen wir uns aber nicht damit beschäftigen wie man ein konkretes Kaskadenmodell erstellt. Stattdessen widmen wir uns der algorithmischen Aufgabe eine geeignete Startmenge S zu wählen. Wir formalisieren die Frage wie folgt: Für eine Startmenge S sei die Wirkungsfunktion $f(S)$ die erwartete Anzahl aktiver Knoten am Ende des Prozesses. Wir nehmen in diesem Abschnitt an, dass die Knotenmenge endlich und damit $f(S)$ durch die Knotenanzahl n beschränkt

ist. Aus wirtschaftlicher Sicht ist $f(S)$ der erwartete Absatz wenn S die Menge der anfänglich gewonnenen Kunden ist. Angenommen uns stünde ein Budget zur Verfügung mit dem k anfängliche Kunden angeworben werden können. Unser Ziel ist es diese k Kunden so auszuwählen, dass der erwartete Absatz maximal ist. Gesucht ist $\max_{|S|=k} f(S)$.

Leider ist für nahezu alle Instanzen der besprochenen Modelle dieses Problem NP-schwierig. Daher ist es das Ziel Unterklassen der Modelle zu finden, die zumindest eine gute Approximation erlauben.
Der Schlüssel zu einer guten Approximation liegt in der Submodularität:
f ist submodular wenn $\forall X \subseteq Y \; \forall v \notin Y : f(X \cup \{v\}) - f(X) \geq f(Y \cup \{v\}) - f(Y)$.
Vergrößert man Y um ein Element v so ist der marginale Zuwachs von f nicht größer als der marginale Zuwachs den v einer Teilmenge X von Y bringt. Submodularität ist eine Art abnehmenden Grenzertrags. Der Nutzen eines weiteren Elements sinkt mit steigender Mächtigkeit der Menge. Ein simple Strategie ist wiederholtes hinzunehmen des Knotens der die maximale marginale Erhöhung bringt. Für eine submodulare und monotone Funktion f kann man mit dieser Strategie auf eine gute Approximation des Optimums hoffen. Intuitiv kann ein scharfes Optimum bei abnehmendem Grenzertrag schwer übersehen werden. Tatsächlich gibt es für diese Strategie eine Leistungsgarantie:

Satz 4.1 (Nemhauser, Wolsey, Fisher 1978)**.** *Sei f monoton und submodular und sei f auf den k-elementigen Mengen bei S^* maximal. Sei S die durch Hill-Climbing gewonnene k-elementige Menge dann ist $f(S^*) \geq f(S) \geq (1 - 1/e)f(S^*) \approx 0.63f(S^*)$*

Glücklicherweise erfüllt das Kaskadenmodell die Bedingungen aus Satz 4.1 .

Satz 4.2 (Kempe, Kleinberg, Tardos 2003,2005)**.** *Für jede Instanz des Kaskadenmodells ist die Wirkungsfunktion f submodular und monoton.*

Beweis. Offensichtlich ist f monoton da der Prozess progressiv ist. Eine alternative Sichtweise auf das unabhängige Kaskadenmodell eröffnet uns einen Weg die Submodularität von f zu zeigen. Ursprünglich hatte ein Knoten wenn er aktiv wurde genau einen Versuch mit Erfolgswahrscheinlichkeit p_{uv} um einen Nachbarknoten v zu aktivieren. Jeder Kante entsprach ein Zufallsversuch. In der alternativen und äquivalenten Sichtweise führen wir diese Versuche für jede Kante *im Voraus* durch. Nunmehr tragen die Kanten keine Erfolgswahrscheinlichkeiten sondern eine 1 für einen Ausbreitungsweg oder eine 0 für keine Übertragung.

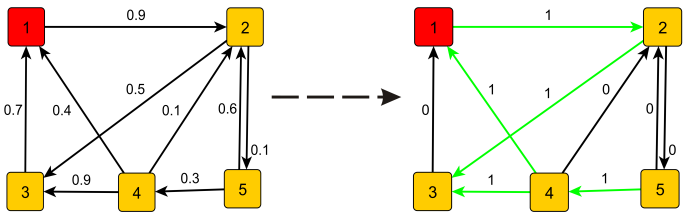

Abbildung 12: Alternative Sichtweise auf das Kaskadenmodell

Im Laufe des Prozesses werden genau die Knoten aktiv, die von der Startmenge aus in einer Einser-Kantenfolge erreicht werden können.
In Abbildung 12 ist das die Menge $\{1, 2, 3\}$. Sei m die Anzahl der Kanten im Graph. Da jede Kante ein Bernoulli-Versuch mit zwei möglichen Ausgängen ist, gibt es 2^m mögliche Ausgänge für die Gesamtheit der Versuche. Sei α einer dieser 2^m Ausgänge und sei $f_\alpha(S)$ die Menge der letztendlich aktiven Knoten bei Startmenge S und Versuchsausgängen α. Sei $R_s^{\langle\alpha\rangle}$ die Menge der Knoten die vom Knoten s aus, bei Ausgangslage α über eine Einser-Kantenfolge erreicht werden können. Dann ist $f_\alpha(S) = |\bigcup_{s\in S} R_s^{\langle\alpha\rangle}|$ die Mächtigkeit einer Vereinigung von Mengen (size-of-union-function) und daher submodular. Schauen wir Abbildung 13 an um einzusehen, dass size-of-union-functions submodular sind. X ist die Vereinigung der blauen Mengen und Y ist X zusammen mit den roten Mengen. Vereinigen wir X und Y mit einer weiteren Menge V dann ist der Zugewinn bei X größer als bei Y. Diese Eigenschaft ist Submodularität.

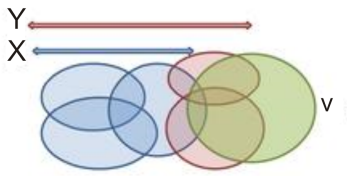

Abbildung 13: Submodularität der size-of-union-function [11]

Wie hängt nun $f_\alpha(S)$ mit $f(S)$ zusammen? Mit einer Wahrscheinlichkeit Prob[α] ist das Ergebnis der Zufallsversuche α. In diesem Fall ist $f_\alpha(S)$ die Anzahl aktiver Knoten am Ende des Prozesses.
Nach Definition ist f ein Erwartungswert und zwar $f(S) = \sum_\alpha \text{Prob}[\alpha] \cdot f_\alpha(S)$. Als Linearkombination submodularer Funktionen mit positiven Koeffizienten ist f ebenfalls submodular. □

5 Eine Empirische Studie

Wie Anfangs erwähnt gab es in den letzten fünfzig Jahren viele empirische Studien zu Diffusionsvorgängen. Ein klassisches Beispiel ist die Studie „Medical Innovations: A Diffusion Study“ von Coleman, Katz und Menzel aus dem Jahr 1966[2]. Die Studie untersucht die Akzeptanz eines neuen Antibiotikums innerhalb der Ärzteschaft von Illinois. Die gute Wirkung des neuen Antibiotikums war den Ärzten aus klinischen Studien und wissenschaftlichen Veröffentlichungen bekannt. Aber die Ärzte begannen erst dann das Antibiotikum zu verschreiben als ihnen Kollegen die Wirkung bestätigten. Die Ausbreitung war im Wesentlichen ein sozialer Prozess. Eine Herausforderung der empirischen Diffusionsforschung ist es, qualitativ und quantitativ zu Erfassen, wie soziale Verbindungen die Diffusion beeinflussen. In der Forschung waren die untersuchten Netzwerke bisher von kleinem Maßstab weil die Datenbeschaffung schwierig war. Dafür wurden diese im Detail untersucht und es wurden hintergründige Einblicke gewonnen. Unsere Modelle

sind qualitativ motiviert durch diese empirischen Studien. Aber die Modelle müssen noch auf quantitativer Ebene an reelle Diffusionsdaten angepasst werden. Allerdings waren die Datensätze aus den traditionellen Studien zu klein um treffende Parameterschätzungen zu machen. Zum Beispiel wie die Annahmewahrscheinlichkeit eines Knotens von der Struktur seiner Nachbarn abhängt. Diese Lücke kann jüngst durch riesige Datensätze aus webbasierten sozialen Netzwerken geschlossen werden.
Im Folgenden betrachten wir eine Studie zum Netzwerk LiveJournal (Backstrom et al. 2006[1]). Livejournal ist eine Webseite auf der Nutzer Blogs erstellen, ein Profil anlegen und für uns besonders wichtig, sich mit Freunden vernetzen und Gruppen beitreten können. Die Webseite hat über eine Millionen aktive Nutzer und mehrere Tausend Gruppen. Aus den Freundeslisten leitet sich das soziale Netzwerk ab und die Mitgliedschaft in einer Gruppe ist das Verhalten dessen Diffusion wir betrachten. Wir untersuchen wie die Wahrscheinlichkeit einer Gruppe beizutreten von der Anzahl der Freunde in dieser Gruppe abhängt. Zu zwei Zeiten (mit einigen Monaten dazwischen) werden die Gruppenzugehörigkeiten verglichen. Für eine Gruppe sei U_k die Menge der Nutzer die zum ersten Zeitpunkt nicht Mitglied waren aber k Freunde in ihr hatten. $P(k)$ sei der Anteil derjenigen aus U_k die dann zur zweiten Zeit Mitglied sind. $P(k)$ ist die empirische Wahrscheinlichkeit zum zweiten Zeitpunkt in einer Gruppe zu sein in der man zum ersten Zeitpunkt nicht war, aber k Freunde darin hatte.

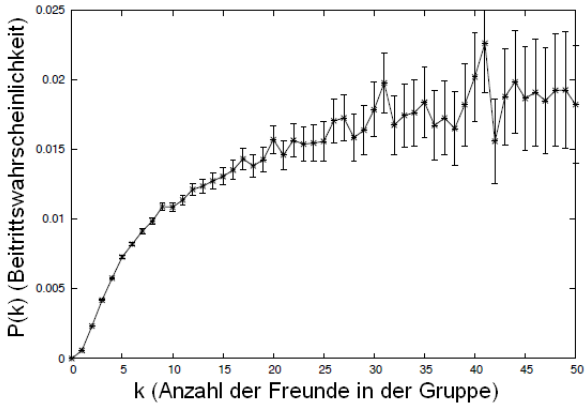

Abbildung 14: Wahrsch. einer Livejournal-Gruppe beizutreten in der man k Freunde hat

Abbildung 14 zeigt $P(k)$ in Abhängigkeit von k für die LiveJournal Daten. Man sieht, dass $P(k)$ von abnehmendem Grenzertrag ist, d.h der Grenzertrag sinkt mit steigendem k. Allerdings gibt es zu Beginn eine Abweichung vom abnehmenden Grenzertrag, denn $P(k)$ steigt von $k = 0$ auf $k = 1$ weniger als von $k = 1$ auf $k = 2$. Mit anderen Worten, einen zweiten Freund in einer Gruppe zu haben gibt der Wahrscheinlichkeit der Gruppe beizutreten einen beträchtlichen Schub, während danach der Effekt des abnehmenden Grenzertrags einsetzt. Es sei darauf hingewiesen, dass der abnehmende Grenzertrag von P nicht mit dem abnehmenden Grenzertrag bzw. der Submodularität von f aus Abschnitt 4 zu verwechseln ist. Eine Funktion $P(k) = \varepsilon \cdot \ln(k)$ wäre ein guter Fit an die Daten. Auch in anderen, neueren Studien mit einem großen Datenumfang wurden ähn-

liche, abnehmende Erträge beobachtet. Zum Beispiel die Wahrscheinlichkeit auf einer Konferenz eine Veröffentlichung zu machen in Abhängigkeit der Zahl der Koautoren die zuvor dort publiziert haben (Backstrom et al. 2006). Es ist eine offene Aufgabe herauszufinden welcher Mechanismus hinter diesem gemeinsamen Effekt steht.

6 Ausblick

Zum Abschluss möchte ich einen Ausblick auf offene Forschungsfragen geben. Die Analyse der LiveJournal Daten ergab auch, dass die Wahrscheinlichkeit zum Gruppenbeitritt abhängig ist von der Vernetztheit der Freunde. Jemand mit k Freunden in einer Gruppe wird mit signifikant größerer Wahrscheinlichkeit beitreten wenn diese k Freunde untereinander viele Kanten haben. Eine verbundene Menge von Freunden hat einen größeren Einfluss als eine vergleichbare Menge voneinander unabhängiger Freunde. In den bisherigen Modellen geht die Struktur des Freundeskreises nicht ein. Eine offene Aufgabe ist es, diese Erkenntnis in ein theoretisches Modell einzubinden. Ein kritischer Punkt ist die Beschaffenheit der Daten. So wurde aus den Freundeslisten bei Livejournal das soziale Netzwerk abgeleitet, doch wissen wir nichts darüber was es bedeutet wenn w als Freund von v gelistet ist. Beide könnten sehr gute Freunde sein oder sich kaum kennen. Wir können auch nicht sagen was einen bestimmten Nutzer dazu motiviert einer Gruppe beizutreten. Zum anderen wurde der Netzwerkzustand nur zu zwei Zeiten betrachtet. Wir wissen nicht was dazwischen passiert ist. Wann ein Individuum (und ob überhaupt) die Gruppenzugehörigkeit seiner Freunde bemerkt hat und wann dies zum Entschluss der Gruppe beizutreten geführt hat. $P(k)$ aus Abbildung 14 macht demnach lediglich eine Aussage über die Gesamtsituation im Netzwerk. Daraus lässt sich nicht ableiten wie ein bestimmtes Individuum reagiert. Die Entwicklung von Online Systemen liefert größere Datenmengen mit besserer Zeitauflösung. Aus diesen werden sich weitergehende Theorien und Modelle entwickeln die zu neuen Erkenntnissen darüber führen werden wie sich Informationen und Verhalten in sozialen Netzwerken verbreiten.

Literatur

[1] L. Backstrom, D. Huttenlocher, J. Kleinberg and X. Lan,*Group formation in large social networks: Membership, growth, and evolution.*, In Proc. 12th Intl. Conf. on Knowledge Discovery and Data Mining, 2006

[2] J. Coleman, E. Katz, and H. Menzel, *Medical Innovations: A Diffusion Study*, Bobbs Merrill, 1966

[3] M. Granovetter, *Threshold models of collective behavior*, Am. J. Sociol., 83:1420-1443, 1978

[4] D. Kempe, J. Kleinberg, and É. Tardos, *Influential nodes in a diffusion model for social networks*, In Proc. 9th Intl. Conf. on Knowledge Discovery and Data Mining, pp. 137-146, 2005

[5] D. Kempe, J. Kleinberg, and É. Tardos, *Influential nodes in a diffusion model for social networks*, In Proc. 32nd Intl. Colloq. on Automata, Languages and Programming, pp. 1127-1138, 2005

[6] S. Morris, *Contagion*, Review of Economic Studies, 67:57-78, 1978

[7] G. Nemhauser, L. Wolsey, and M. Fisher, *An analysis of the aproximations for maximizing submodular set functions*, Math. Programm, 14:265-294, 1987

[8] E. Rogers, *Diffusion of innovations*, 4th ed. Free Press, 1995

[9] T. Schelling, *Micromotives and Macrobehavior*, Norton, 1978

[10] D. Strang and S. Soule, *Diffusion in organizations and social movements: From hybrid corn to poison pills*, Ann. Rev. Sociol., 24:265-290, 1998

[11] J. Leskovec, *Diffusion and Cascading Behavior in Social Networks*, Tutorial at NATO Advanced Study Institute on Mining Massive Data Sets for Security, 2007 Folien