Inferenzstatistische Modellierung der Dynamik bipartiter Netzwerke am Beispiel einer online Reiseplattform

Roman Tilly Johannes Putzke David Schölgens Kai Fischbach
Department of Information Systems and Information Management, University of Cologne
Pohligstr. 1, 50969 Köln, Germany
{tilly; putzke; schoelgens; fischbach}@wim.uni-koeln.de

ZUSAMMENFASSUNG

Ziel dieses Beitrages ist die Vorstellung einer inferenzstatistischen Methode zur Modellierung der Dynamik bipartiter Netzwerke. Exemplarisch wird die Methode an einem bipartiten Netzwerk aus Reisezielen und Benutzern einer online Reiseplattform illustriert. Unser akteursbasiertes Modell untersucht dabei Faktoren, die Einfluss darauf haben, zu welchem Reiseziel ein Benutzer der Reiseplattform (wie z. B. `www.tripadvisor.com`) einen Reisebericht schreibt. Für den Zeitraum von 2006 bis 2009 wurden mehrere bipartite Netzwerke modelliert, deren Knoten durch Benutzer und Reiseziele und deren Kanten durch Reiseberichte repräsentiert wurden. Dieser Ansatz ist, nach unserer Kenntnis, die erste inferenzstatistische Modellierung der Dynamik eines bipartiten Netzwerks zur Untersuchung des Reiseverhaltens von Akteuren. Er kann von Wissenschaftlern und Unternehmen weiterentwickelt werden, um Reiseströme vorherzusagen. Da die Modellierung der Dynamik von (bi-)partiten Netzwerken für eine Vielzahl von Fragestellungen in der Wirtschaftsinformatik Relevanz hat, liegt der Schwerpunkt dieses Artikels weniger auf der inhaltlichen Interpretation der Ergebnisse als auf der grundlegenden Darstellung der Modellklasse.

Schlüsselwörter

Akteursbasierte Modellierung; Reise; Tourismus; e-Tourism; Netzwerkanalyse; SIENA; ERGM

1. EINFÜHRUNG

Reiseplattformen im Internet werden immer häufiger für die Informationssuche und Buchung von Reisen genutzt. Die Zahl der Benutzer dieser Plattformen und der von ihnen verfassten Reiseberichte wächst stetig (s. Abbildung 1 für die untersuchte Internetplattform). Unter online Reiseplattformen verstehen wir Internetseiten wie z. B. `www.igougo.com`, `www.tripadvisor.com` oder `www.holidaywatchdog.com`, die es Benutzern ermöglichen, Reiseberichte anderer Benutzer über verschiedene Reiseziele zu lesen und eigene Berichte zu verfassen. Darüber hinaus können die Benutzer oft auch eine persönliche Profilseite pflegen, mit anderen Benutzern mittels privater Nachrichten kommunizieren und die Reiseberichte anderer Benutzer kommentieren. Der Großteil der Daten dieser Reiseplattformen ist im Internet frei zugänglich.

Ziel dieses Artikels ist es, einen Ansatz vorzustellen, um solche Faktoren zu finden, die die Benutzer der Reiseplattform in der Wahl eines Reiseziels beeinflussen. Dazu haben wir Daten einer Reiseplattform gesammelt, die in Abschnitt 3 dargestellt sind. Das Reiseverhalten der Benutzer haben wir mit der in Abschnitt 4 beschriebenen akteursbasierten Modellierung untersucht. Benutzer und Reiseziele wurden als Akteure in einem Netzwerk modelliert und Reisen eines Benutzers zu einem Reiseziel – ausgedrückt durch einen Reisebericht auf der Internetseite – als Verbindung zwischen beiden. Die Ergebnisse der Analyse der Einflussfaktoren für den Auf- und Abbau einer derartigen Verbindung werden in Abschnitt 5 dargestellt, ehe die Ergebnisse in Abschnitt 6 interpretiert und kritisch beleuchtet werden sowie deren Relevanz für Forschung und Praxis herausgearbeitet wird.

Unser Ansatz kann von Wissenschaftlern und Unternehmen der Tourismus- bzw. Reisebranche genutzt und weiterentwickelt werden, um Einflussfaktoren zu identifizieren, die für Benutzer der Internetplattform gelten. Sie lassen sich unter Umständen auch über das Internet hinaus nutzen. Wie [15] gezeigt haben, erlaubt die Analyse von Daten aus virtuellen Welten sinnvolle Rückschlüsse auf die reale Welt. Im Falle von Reiseplattformen könnte dies am Ende zu besseren Vorhersagen von Reisetrends führen.

Darüber hinaus gehen wir davon aus, dass sich der Ansatz grundsätzlich auch auf andere Domänen als die Tourismusbranche übertragen lässt, in denen Benutzer bzw. Kunden wiederholt zwischen verschiedenen Produkten oder Diensten wählen und ihre Wahl durch die Veröffentlichung eines Erfahrungsberichts offenlegen. So ließe sich der Ansatz mit geringer Anpassung auch auf Internetseiten wie z. B. `www.ciao.com` oder `www.qype.com` übertragen.

2. EINORDNUNG DES ARTIKELS

Für die Vorhersage von Reiseströmen oder -trends wurden in den letzten 45 Jahren verschiedene Methoden entwickelt. Einen Überblick über quantitative Methoden wie autoregressive Zeitreihenmodelle und ökonometrische Modelle geben [26] (Arbeiten bis 1995) und [22] (Arbeiten von 2000 bis 2006). Diese Methoden verwenden als abhängige Variable meist die Anzahl der Touristen aus einem Herkunftsland in einem Zielland oder die Ausgaben der Touristen aus einem

10^{th} International Conference on Wirtschaftsinformatik,
$16^{th} - 18^{th}$ February 2011, Zurich, Switzerland

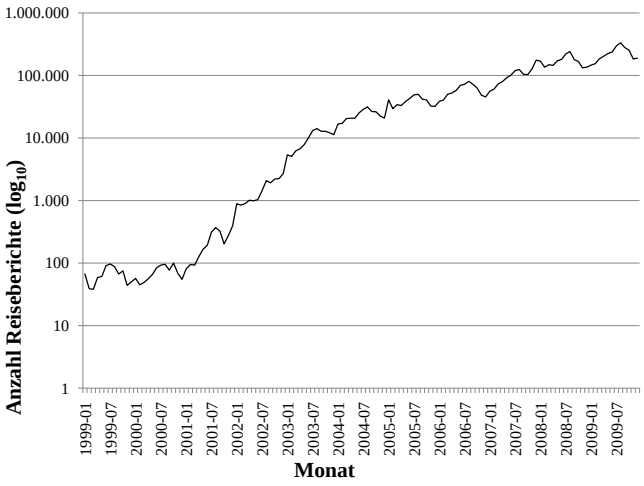

Abbildung 1: Entwicklung der Reiseberichte pro Monat für die untersuchte Internetplattform, Kriterium: Monat der Reise wie im Reisebericht angegeben

Herkunftsland im Zielland.[1] Andere Größen sind, je nach Zielsetzung der Analyse, die Anzahl verbrachter Nächte am Ziel oder der Marktanteil eines bestimmten Zieles. Als erklärende Größen werden bspw. häufig die Bevölkerungszahl oder das durchschnittliche Einkommen im Herkunftsland, die Kosten für die Reise zum Ziel oder die Lebenshaltungskosten am Ziel verwendet.

Im Vergleich dazu nutzt nur eine beschränkte Anzahl von Arbeiten Daten, die von Kunden durch ihre Aktivitäten auf Reiseplattformen im Internet hinterlassen werden (z. B. [5, 10, 17, 12, 16, 13, 11]). Die meisten dieser Arbeiten verfolgen dabei jedoch einen explorativen oder Datamininggetriebenen Ansatz und vernachlässigen traditionelle Modellierungsansätze der Tourismusforschung, wie sie sich etwa in Form von Choice-Modellen manifestiert hat (vgl. etwa [1]).

An dieser Stelle setzen wir an. Unser Artikel versucht, traditionelle Choice-Modelle der Tourismusforschung auf die von Anwendern auf Reiseplattformen im Internet hinterlassenen Daten anwendbar zu machen. Hierbei kombinieren wir Choice-Modelle, Wartezeitmodelle und Netzwerkmodellierung in einem ganzheitlichen Ansatz. Die vorgeschlagene Methodik ist nach unserem Wissen damit die erste Anwendung einer inferenzstatistischen Methode auf die Dynamik eines bipartiten Netzwerks zur Untersuchung des Reiseverhaltens von Akteuren.

Unser Modell ist der Klasse der akteursbasierten Modelle und der Familie der exponentiellen Zufallsgraphenmodelle zuzurechnen[2]. Zu dieser Familie von Modellen gehören loglineare Modelle wie das bekannte p_1-Modell [7] und verschiedene Erweiterungen, aber auch die Modelle auf Basis zeitkontinuierlicher Markow-Prozesse[3].

3. DATENBASIS

[1]Diese und die folgenden Größen nach [26].

[2]Ausführliche Erläuterungen hierzu finden sich in [4, 14, 25].

[3]Auch Markow-Modelle genannt; Grundlagen in diesem Bereich legten [4] und [6].

Entitätstyp	Anzahl Entitäten
Reiseziele	60.369
Freizeitangebote	1.139.003
Foren	15.177
Diskussionsthemen	2.945.947
Diskussionsbeiträge	18.895.497
Reiseberichte	7.885.482
Benutzerprofile	3.874.768

Tabelle 1: Absolute Menge der gesammelten Daten

In diesem Abschnitt stellen wir die Daten vor, die für die Analyse der Einflussfaktoren auf das Reiseverhalten verwendet wurden. Grundlage hierfür waren Daten einer großen Internetplattform für Reiseberichte, die im Zeitraum von Oktober 2009 bis März 2010 mit Hilfe eines Web-Crawlers erhoben wurden. Dieser verarbeitete systematisch alle auffindbaren Internetseiten der Plattform und extrahierte die relevanten Informationen. Die Daten umfassen Entitäten der folgenden Typen:

Reiseziele: Geografische Orte oder Regionen, z. B. Städte, Bundesländer, Staaten.

Freizeitangebote: Angebote oder Attraktionen verschiedenen Typs, z. B. Hotels, Restaurants, Vergnügungsparks, Museen, die an einem Reiseziel gelegen sind.

Foren: Diskussionsplattformen auf der Internetseite, die jeweils einem Reiseziel zugeordnet sind.

Diskussionsthemen: Austausch mehrerer Benutzer in Foren.

Diskussionsbeiträge: Von Benutzern verfasste Beiträge zu einzelnen Diskussionsthemen in Foren.

Reiseberichte: Von Benutzern verfasste Erfahrungsberichte zu Freizeitangeboten einschließlich deren Bewertung auf einer fünfstufigen Punkteskala.

Benutzerprofile: Persönliche Seite jedes Benutzers mit demografischen Angaben und Beschreibung der individuellen Reisepräferenzen.

Verglichen mit den Größenangaben, die die Betreiber der Plattform auf der Internetseite veröffentlichen, konnten schätzungsweise über 85 Prozent der verfügbaren Daten gesammelt werden. Die absoluten Zahlen der Entitäten je Typ sind in Tabelle 1 aufgeführt.

Nach der Vorstellung der Datengrundlage erläutern wir im folgenden Abschnitt die akteursbasierte Modellierung dieser Daten.

4. MODELL

Die Daten der Reiseplattform wurden als bipartites Netzwerk bzw. bipartiter Graph modelliert.[4] Ein Graph besteht aus *Knoten* und *Kanten*, wobei einzelne Knoten durch Kanten verbunden sein können. Formal ist ein Graph ein Paar $G = (V, E)$ disjunkter Mengen mit $E \subseteq [V]^2$, wobei V die Knotenmenge ($n = |V|$) und E die Kantenmenge ($l = |E|$) bezeichnet [3]. Graphen können gerichtet oder ungerichtet sein. Im *gerichteten* Fall haben alle Kanten einen Ursprungs- und einen Endknoten, im *ungerichteten* Fall wird nicht zwischen Ursprungs- und Endknoten unterschieden.

Ein *bipartiter* Graphen ist wie folgt definiert: „Es sei $r \geq 2$ eine natürliche Zahl. Ein Graph $G = (V, E)$ heißt *r-partit*,

[4]Die Begriffe Netzwerk und Graph verwenden wir in diesem Artikel synonym. In anderen Kontexten ist eine differenzierte Betrachtung erforderlich, wie sie bspw. in [24] vorgenommen wird.

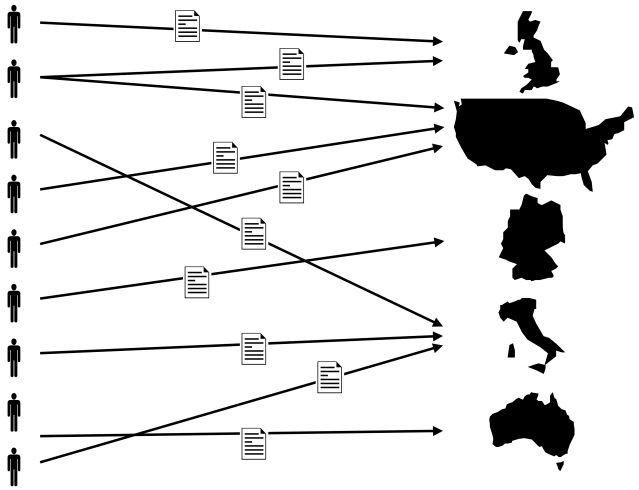

Abbildung 2: Beispielhafte Darstellung des Benutzer-Reiseziel-Netzwerks

wenn eine Partition von V in r Teile existiert, so dass die Endecken[5] einer jeden Kante von G in verschiedenen Partitionsklassen liegen: Ecken aus der gleichen Klasse dürfen nicht benachbart[6] sein. Ein 2-partiter Graph heißt auch *bipartit* (oder *paar*).“[3]
Die Knotenmenge V eines r-partiten Graphen zerfällt dann in $V_1 \dots V_q \dots V_r$ mit $n_q = |V_q|$ ($\bigcup_{q=1}^{r} V_q = V$). Für die Analyse der Reiseplattform wurden Benutzer als eine Menge V_1 von Knoten modelliert, Reiseziele als eine weitere Knotenmenge V_2. Ein Reisebericht eines Benutzers $v_a \in V_1$ zu einem Reiseziel $v_b \in V_2$ wurde im Graph als Kante zwischen den Knoten v_a und v_b modelliert.[7] Diese Netzwerkmodellierung ist beispielhaft in Abbildung 2 dargestellt.

Für den Zeitraum von Anfang 2006 bis Ende 2009 wurden die Reiseberichte eines Kalenderjahres zu einem Beobachtungszeitpunkt B zusammengefasst und in einer Adjazenzmatrix $x(B)$ dargestellt, um die Entwicklung des Netzwerks im Laufe der Zeit untersuchen zu können (longitudinale Analyse). Die Adjazenzmatrix ist eine $n \times n$-Matrix, die die Konfiguration $G(t)$ eines Graphen G zum Zeitpunkt t darstellt. Als Konfiguration bezeichnet man die vollständige Beschreibung, zwischen welchen Knoten eines Graphen zum Zeitpunkt t Kanten existieren und zwischen welchen nicht. Einzelne Einträge in der Adjazenzmatrix werden durch x_{ij} bezeichnet. Besitzt der Knoten i zum Zeitpunkt t eine Kante zum Knoten j, so definieren wir $x_{ij}(t) = 1$, andernfalls $x_{ij}(t) = 0$.
Die Umwandlung von Ereignisdaten in Zustandsdaten wurde in der einschlägigen Literatur eingehend diskutiert (vgl. etwa [23]). In diesem Fall wurde eine jahresweise Aggregation der Reiseberichte vor allem aus zwei Gründen vorgenommen. Zum einen findet der Zeitpunkt der Reise und die Erstellung des Reiseberichte durch die meisten Nutzer nicht zeitgleich statt, sondern diese Zeitspanne variiert zwischen Benutzern und Reiseberichten. Überdies hat die Aggregation den Vorteil, dass die nicht unerheblichen saisonalen Schwankungen innerhalb eines Jahres ausgeglichen werden. So variieren etwa die relativen Anteile einzelner Monate eines Jahres an den gesamten Reiseberichten in diesem Jahr um $4,43 - 11,89$ Prozentpunkte für die Jahre 2000 bis 2009.

[5]Ecke: andere Bezeichnung für Knoten, Anm. der Autoren.
[6]Zwei Knoten sind genau dann benachbart, wenn eine Kante existiert, die beide Knoten direkt verbindet, Anm. der Autoren.
[7]Reiseberichte beziehen sich zunächst nicht direkt auf geografische Reiseziele, sondern auf Freizeitangebote, die jeweils an einem Reiseziel gelegen sind. Für die Analyse wurden die Reiseberichte jedoch nach Reisezielen aggregiert.

4.1 Die Klasse akteursbasierter Modelle

Die in [18, 19, 20] beschriebene Klasse der akteursbasierten Modelle ermöglicht es, die vorherzusagende Anzahl der Reiseberichte (als Proxy für die Anzahl der Reisen an einen bestimmten Zielort), Netzwerkvariablen und Charakteristiken der Akteure gleichzeitig als abhängige und unabhängige Variablen zu modellieren und deren gegenseitigen Einfluss aufeinander zu quantifizieren. Für die Anwendung bei der Analyse der Reiseplattform ist daher wichtig, dass sie zum einen die longitudinale Analyse der Entwicklung des Netzwerks über mehrere Beobachtungszeitpunkte erlaubt, was ein Vorteil gegenüber anderen, statischen Modellen ist. Zum anderen erlaubt es diese Modellklasse, bipartite Graphen in ihrer Ursprungsform zu analysieren. Andere Modelle lassen nur nicht-partite Graphen zu. Im Prinzip lässt sich zwar jeder bipartite Graph auf einen nicht-partiten Graphen projizieren. Das hat jedoch einige Nachteile [9]. Die Modellklasse soll hier kurz vorgestellt werden. Für eine detaillierte Einführung sei auf die genannten Quellen verwiesen.

Die akteursbasierten Modelle modellieren das Entscheidungskalkül einzelner Akteure, die zwischen einer endlichen Menge von Alternativen wählen können. Die Akteure sind Teil eines Netzwerks mehrerer Akteure und die Alternativen bestehen darin, eine Verbindung zu einem anderen Akteur aufzubauen, eine bestehende Verbindung aufzulösen oder nichts zu tun. Dieser Zusammenhang wird durch einen Graphen modelliert, wobei Akteure durch Knoten und Verbindungen durch Kanten repräsentiert werden. Jeder Akteur kann selber darüber entscheiden, zu welchem anderen Akteur er eine Verbindung aufbaut. Der Zustand aller Verbindungen entspricht der Konfiguration $G(t)$ des Graphen und lässt sich durch die Adjazenzmatrix $x(t)$ darstellen. Wenn X die Menge der Adjazenzmatrizen aller möglichen Konfigurationen ist, kann man die oben genannte Entscheidung formal als sogenannten *Mikroschritt* von einer Konfiguration bzw. Adjazenzmatrix $x(t) \in X$ zu einer der möglichen Konfigurationen bzw. Adjazenzmatrizen $x' \in X$ auffassen. Für einen solchen Mikroschritt gilt zusätzlich die Bedingung, dass sich $x(t+1)$ von $x(t)$ in höchstens einem Eintrag x_{ij} unterscheiden darf (aufgrund der zuvor genannten Alternativen Verbindung aufbauen / auflösen oder nichts tun). Für die Längsschnittanalyse eines Netzwerks müssen dessen Graphkonfigurationen bzw. Adjazenzmatrixdarstellungen für mehrere Beobachtungszeitpunkte vorliegen. In der Klasse der akteursbasierten Modelle wird angenommen, dass die Unterschiede der Konfigurationen von einem Beobachtungszeitpunkt zum nächsten die Folge einzelner Mikroschritte sind, die zwischen zwei Beobachtungszeitpunkten stattfinden. Durch die Erklärung der Mikroschritte lässt sich dann die Gesamtveränderung des Netzwerks erklären.
Die Zeitpunkte, zu denen ein Akteur $i \in V$ die Möglichkeit für einen Mikroschritt hat, werden in einem akteursbasierten Modell durch die *Ratenfunktion* $\lambda_i(x)$ beschrieben. Sie folgt

einem Poisson-Prozess, wodurch die Zeitabstände zwischen zwei Mikroschritten exponentialverteilt sind. Die Parameter der Verteilung werden aus den Daten geschätzt.
Weitere Annahmen der Modellklasse sind, dass zu jedem Zeitpunkt nur die aktuelle Konfiguration des Netzwerks (probabilistisch) die weitere Entwicklung bestimmt (Markow-Prozess), dass die Akteure selber über ihre ein- und ausgehenden Kanten entscheiden – was impliziert, dass sie das gesamte Netzwerk wahrnehmen können und ihre Wahlentscheidung hierüber optimieren – und, dass Veränderungen im Netzwerk immer nacheinander stattfinden, d. h., dass es insbesondere keine koordinierten, zeitgleichen Veränderungen durch mehrere Akteure gibt.

Die Präferenz eines Akteurs i für eine einzelne Konfiguration $x \in X$ wird durch die *Zielfunktion*[8]

$$f_i(\beta, x) = \sum_{k=1}^{K} \beta_k s_{ik}(x)$$

beschrieben.[9] Die $s_{ik}(x)$ sind K sogenannte *Effekte*: Funktionen von Graph und Akteur/en, die sich zum Zielfunktionswert addieren und durch Theorie sowie domänenspezifisches Wissen motiviert sind. Sie stellen bestimmte Kenngrößen einer Konfiguration x und / oder der Eigenschaften eines oder mehrerer Akteure[10] aus Sicht des Akteurs i dar und es wird vermutet, dass diese Kenngrößen einen (positiven oder negativen) Einfluss auf die Bewertung einer Konfiguration haben. Die Gewichte β_k repräsentieren jeweils die Stärke des Einflusses des k-ten Effekts auf die Gesamtpräferenz. Sie sind die statistischen Modellparameter, die später geschätzt werden und die je nach Größe und Signifikanz eine Aussage darüber zulassen, ob ein Effekt einen nennenswerten Einfluss darauf hat, ob ein Akteur eine Verbindung aufbaut oder auflöst.
In der Klasse der akteursbasierten Modelle sind verschiedene generische Effekte definiert, die je nach Kontext ausgewählt und interpretiert werden können. Sie sind, wie auch die gesamte Zielfunktion, aus der Perspektive eines individuellen Akteurs zu verstehen. Grundsätzlich werden zwei verschiedene Arten von Effekten unterschieden: *Strukturelle Effekte* berechnen sich ausschließlich aus vorhandenen bzw. fehlenden Kanten. *Kovariateneffekte* berechnen sich aus strukturellen Eigenschaften des Netzwerks und Eigenschaften von Akteuren bzw. Verbindungen zwischen Akteuren. Kovariaten werden wiederum zum einen danach unterschieden, ob sie *monadisch* sind, d. h. sich auf Eigenschaften von einem oder zwei Akteur/en beziehen, oder ob sie *dyadisch* sind, d. h. Informationen über die Verbindung zwischen zwei Akteuren geben. Zum anderen wird unterschieden, ob die Kovariaten über alle Beobachtungszeitpunkte *konstant* sind oder *veränderlich*. Des Weiteren werden die monadischen Kovariateneffekte danach unterschieden, ob sie sich auf Eigenschaften des Akteurs i selbst (*ego*-Effekt), auf Eigenschaften anderer Akteure (*alter*-Effekt) oder auf beide (*ego* × *alter*-Effekt) beziehen.

[8]Zu der Zielfunktion wird im Modell noch eine unabhängig normalverteilte Zufallsvariable addiert, die den Restfehler (Residuum) der unerklärten Präferenz darstellt. Sie wird in diesem Artikel nicht weiter explizit behandelt.
[9]Diese und die folgenden Erläuterungen dieses Abschnitts nach [19, 21, 15].
[10]Einschließlich des Akteurs i selbst.

Benutzer	Anzahl Reiseberichte
Partnerseite 1	524.935
Partnerseite 2	402.520
Partnerseite 3	115.912
Partnerseite 4	91.482
Partnerseite 5	52.802
Partnerseite 6	39.682
Partnerseite 7	24.651
Partnerseite 8	21.540
Partnerseite 9	17.735
Partnerseite 10	2.961
Summe Partnerseiten	1.294.220
Gelöschte Benutzer	569.970
SUMME	1.864.190

Tabelle 2: Anzahl der Berichte aller Partnerseiten und ehemaligen Benutzer

Die Wahrscheinlichkeit dafür, dass sich ein Akteur i zum Zeitpunkt t für eine bestimmte Konfiguration $x \in X$ entscheidet, ist durch

$$\frac{\exp(f_i(\beta, x))}{\sum_{x' \in X} \exp(f_i(\beta, x'))}$$

gegeben [19, 21]. Um den Einfluss verschiedener Faktoren auf die Entscheidungen der Akteure zu quantifizieren, werden die Gewichte β_k mittels logistischer Regression geschätzt.

4.2 Bildung der Netzwerke für die Analyse

Für die Untersuchung haben wir den Datensatz der Reiseplattform eingeschränkt und kleinere Netzwerke aus Akteuren und Reisezielen gebildet. Die Gründe und Vorgehensweise hierfür werden im Folgenden erläutert.
Knapp ein Viertel (23, 64 Prozent) der Reiseberichte im Datensatz lassen sich nicht (mehr) einzelnen Benutzern zuordnen. Sie stammen entweder von Partnerseiten der Plattform oder wurden von Benutzern verfasst, die ihre Mitgliedschaft inzwischen beendet haben. Berichte von Partnerseiten wurden auf einer anderen Internetseite als der Plattform verfasst, von dem Betreiber aber an diese weitergegeben. Als Verfasser ist bei diesen Berichten kein einzelner Benutzer angegeben, sondern jeweils ein Platzhalter für alle Berichte von einer Partnerseite. Im Gesamtdatensatz wurden Berichte von zehn solcher Partnerseiten gefunden, die in Tabelle 2 anonymisiert mit der jeweiligen Anzahl an Reiseberichten aufgeführt sind. Alle Berichte von ehemaligen Benutzern tragen einen immer gleichen Platzhalter als Verfasser, so dass diesem Platzhalter auch zahlreiche Berichte zuzuschreiben sind, die ebenfalls in Tabelle 2 aufgeführt sind.
Da sich Reiseberichte und Diskussionsbeiträge dieser Benutzer nicht einzelnen, realen Menschen zuordnen lassen, sondern Aggregationen größerer Gruppen von Reisenden darstellen, ist es nicht sinnvoll, diese Benutzer als Akteure im Netzwerk zuzulassen.

Eine wichtige Annahme der akteursbasierten Modelle ist, dass alle Akteure jederzeit das gesamte Netzwerk beobachten können, um bei ihren Entscheidungen über den Auf- oder Abbau von Verbindungen zwischen allen Optionen wählen zu können und die Entscheidungen anderer Akteure im Blick zu haben. Diese Annahme ist bei einem Netzwerk, das so groß ist wie das der untersuchten Reiseplattform, zu hinterfragen.
Der Fokus unserer Untersuchung liegt jedoch weniger auf der

Repräsentativit der Ergebnisse für alle Benutzer der Reiseplattform, als vielmehr auf der internen Validität der Ergebnisse und der grundsätzlichen Beurteilung der akteursbasierten Modellierung zur Erklärung der Reiseentscheidungen. Aus diesem Grund haben wir die Menge der Benutzer auf solche eingeschränkt, die auf der Reiseplattform sehr aktiv sind. Neben einigen anderen Kriterien hieß das vor allem, dass sie sich dadurch auszeichneten, dass sie die meisten Reiseberichte verfasst und so ihr Reiseverhalten am stärksten offengelegt hatten. Diese Einschränkung erhöht die empirische Genauigkeit der Analyse. Für diese sehr aktiven Benutzer sehen wir außerdem die genannte Modellannahme (Sichtbarkeit des gesamten Netzwerks) als erfüllt an, da auch [2] zeigt, dass sehr aktive Nutzer in (Online-) Netzwerken sich gegenseitig kennen und beobachten. Dennoch bildeten wir anhand der im folgendenden genannten Kriterien fünf verschiedene Subnetzwerke der aktivsten Nutzer, um auch die externe Validität unserer Analyse sicherzustellen.
Bei der Einschränkung der Nutzerbasis wurden zunächst nur solche Benutzer zugelassen, die auf ihrer Profilseite mindestens Alter, Geschlecht und Herkunft angegeben und die mindestens einen Reisebericht und mindestens einen Diskussionsbeitrag geschrieben hatten. Bei diesen Benutzern gehen wir davon aus, dass sie ein Mindestmaß an Aktivität auf der Internetplattform zeigen, indem sie ihre Profilseite zumindest rudimentär pflegen und die Funktionen des Verfassens von Reiseberichten und Diskussionsbeiträgen grundsätzlich nutzen.
In einem weiteren Schritt wurden anhand unterschiedlicher Filterkriterien für Benutzer und Reiseziele verschiedene Subnetzwerke gebildet. Für die ersten beiden Subnetzwerke wurden die 100 Benutzer ausgewählt, die in den Jahren 2007 bis 2009 die meisten Berichte zu mindestens 15 (Subnetzwerk 1) bzw. 30 (Subnetzwerk 2) verschiedenen Reisezielen geschrieben hatten. Diese Kriterien stellen sicher, dass wir erstens sehr aktive Benutzer für die Analyse erhalten, was für die Zielsetzung der hohen internen Validität von Bedeutung ist. Zweitens gilt durch die Vorgabe, dass die Reiseberichte zu *verschiedenen* Reisezielen erfolgen mussten, für diese Benutzer, dass sie bei ihren Reiseentscheidungen viele mögliche Reiseziele in Betracht ziehen. Das wiederum stützt die Modellannahme, dass die Benutzer bei Ihren Entscheidungen das gesamte Netzwerk beobachten können. Dabei liegt bei dem Filterwert von 30 in etwa die Obergrenze, bei der überhaupt noch Benutzer alle Filterkriterien (inklusive des folgenden) erfüllen.
Die gleichen Anforderungen an die Benutzer galten analog für die Subnetzwerke 3 und 4, wobei die Benutzer zusätzlich zu den Top-1.000 Verfassern von Diskussionsbeiträgen gehören mussten. Dieses zusätzliche Kriterium wurde eingeführt, um zu analysieren, ob Benutzer, die sich zusätzlich zu Reiseberichten auch stark am interaktiven Austausch untereinander in Diskussionen beteiligen und so informieren, in ihren Entscheidungen anders beeinflusst werden. Für das fünfte Subnetzwerk galten für die Benutzer die gleichen Filterkriterien wie für Subnetzwerk 4 (Berichte zu mindestens 30 verschiedenen Reisezielen von 2007 bis 2009, Top-100-Verfasser von Berichten, Top-1.000-Verfasser von Diskussionsbeiträgen).
Die Reiseziele wurden für die einzelnen Subnetzwerke entweder auf die Ebene von Bundesländern oder von Städten (bzw. Landkreisen) aggregiert und geografisch außerdem auf bestimmte Staaten oder Kontinente beschränkt. So aggregiert bzw. begrenzt kann man davon ausgehen, dass auch die Menge der Reiseziele im Subnetzwerk für die Benutzer beobachtbar ist. Für die Subnetzwerke 1 bis 4 wurden die Reiseziele auf Bundesländer der USA begrenzt und aggregiert. Für das fünfte Subnetzwerk wurde die Filterung auf Bundesländer in Nordamerika erweitert.
Die USA wurden ausgewählt, da sie das Land sind, das mit Abstand die meisten Reiseberichten (35, 594%; im Vergleich dazu Großbritannien am zweit meisten mit 8, 554%) erhalten hat. Für das fünfte Subnetzwerk wurde mit Nordamerika der Kontinent mit den meisten Reiseberichten (38, 494%; Europa auf Platz zwei mit 35, 841%), um zu prüfen, ob dies zu Veränderungen in den Entscheidungen der Benutzer führt.
Auch diese Einschränkungen auf das Land bzw. den Kontinent mit den meisten Reiseberichten folgen der Prämisse der hohen Genauigkeit und internen Validität der Daten.

Alle fünf vorgestellten Subnetzwerke erfüllten überdies die Modellannahmen eines Jaccard-Indices nicht kleiner als $0,2$ [21]. Der Jaccard-Index [8] beschreibt allgemein die Ähnlichkeit zweier Mengen und in der Anwendung auf Netzwerke die Kontinuität eines Netzwerks von einem Beobachtungszeitpunkt zum nächsten. Er ist als

$$\frac{N_{11}}{N_{11} + N_{01} + N_{10}}$$

definiert. Dabei ist N_{11} die Anzahl der Verbindungen, die in beiden Netzwerken vorhanden sind, N_{01} die Anzahl der neu hinzugekommenen Verbindungen und analog N_{10} die Anzahl der aufgelösten Verbindungen.
Für die Anwendung eines akteursbasierten Modells sollte die Veränderung des Netzwerks von einem Beobachtungszeitpunkt zum nächsten nicht zu groß sein (Jaccard-Index optimalerweise $> 0,3$, aber nicht $< 0,2$), damit die Veränderung weiterhin als Markow-Prozess interpretiert werden kann [21].

4.3 Schätzverfahren

Für die Schätzung der Regressionskoeffizienten wurde das R-Paket RSiena (Version 1.0.11, Revision 84) verwendet.

Wie in allen statistischen Modellen besteht die Schwierigkeit darin, dass ein Effekt die Koeffizienten anderer Effekte verändern kann, je nachdem ob er einbezogen wird oder nicht [21]. Gleichzeitig wird aber die Schätzung der Koeffizienten umso zeitaufwendiger und der Schätzalgorithmus unter Umständen instabil, je komplexer das Modell ist, das heißt, je mehr Effekte einbezogen werden. Aus diesen Gründen ist weder ein Verfahren sinnvoll, in dem zu einem Grundmodell mit wenigen Effekten nur Effekte hinzugefügt werden, noch eines, in dem aus einem Gesamtmodell mit allen Effekten sukzessive Effekte ausgeschlossen werden. Vor diesem Hintergrund wurden in einem mehrstufigen Verfahren verschiedene konkrete Modelle, d. h. unterschiedliche Kombinationen der insgesamt 16 Effekte (s. Abschnitt 5), gebildet und die Koeffizienten geschätzt.
Im ersten Schritt wurden alle Effekte in ein Gesamtmodell integriert und die Koeffizienten geschätzt, um einen ersten Eindruck der Stärke der einzelnen Effekte zu bekommen.[11]
Anschließend wurden sechs Modelle mit bestimmten Kombinationen von Effekten geschätzt, da bei 16 Effekten nicht

[11]Bei allen hier berichteten Modellen konvergierte die Schätzung zu einem Niveau $t < 0,1$.

alle möglichen Kombinationen der Effekte sinnvoll sind. Die Effekte wurden dabei so kombiniert, dass das Potenzial für Korrelationen zwischen den Effekten gering war.[12]
Im dritten Schritt wurde zunächst ein Basismodell mit den Effekten geschätzt, die in den vorigen Modellen die größten Werte für die (approximativ normalverteilte) Teststatistik $\hat{\mu}/\hat{\sigma}$ [18] aufgewiesen hatten. In das Basismodell wurden nur diejenigen Effekte aufgenommen, die sich in allen Modellschätzungen der ersten beiden Schritte als statistisch signifikant erwiesen hatten ($\alpha < 0,05$). Das Basismodell wurde anschließend mit Kombinationen von Effekten erweitert, die zumindest in einigen Modellen der ersten beiden Schritte signifikant zum 5-Prozent-Niveau waren. Bei der Kombination dieser Effekte wurden die Korrelationen zwischen den Effekten berücksichtigt, die in den ersten beiden Schritten aufgetreten waren. Effekte, die miteinander korrelierten ($\rho > 0,2$), wurden nach Möglichkeit nicht miteinander kombiniert.
Im letzten Schritt wurden auf Grundlage der Ergebnisse der vorigen Schritte mehrere finale Modelle für das Netzwerk gebildet, die auch im folgenden Beitrag berichtet werden sollen. Hierfür wurden nur Effekte ausgewählt, die im dritten Schritt immer ein Signifikanzniveau von $\alpha < 0,05$ erfüllt hatten. Wenn ein Effekt dieses Niveau in einem der Modelle nicht erreicht hatte, wurde er im letzten Schritt dennoch berücksichtigt, sofern die niedrigere Signifikanz nur auf Korrelationen mit anderen Effekten zurückzuführen war, die im finalen Modell nicht mehr enthalten waren. Kombinationen von Effekten, die im dritten Schritt einen Korrelationskoeffizienten absolut größer $0,2$ aufgewiesen hatten, wurden im letzten Schritt nach Möglichkeit nicht kombiniert. In einem solchen Fall wurde je ein finales Modell mit jedem der untereinander korrelierten Effekte gebildet.

5. ERGEBNISSE

Die Ergebnisse der finalen Modelle aller Subnetzwerke sind in den Tabellen 3 (ausführlich für die Subnetzwerke 1 und 2) und 4 (zusammengefasst für die Subnetzwerke 3 bis 5) dargestellt. Die einzelnen Effekte werden im Folgenden näher erläutert. Wenn die Daten für einen Effekt logarithmiert wurden (Basis e), um sie in der Größenordnung den Daten der anderen Effekte anzupassen, ist dies jeweils angegeben.

Strukturelle Effekte

Dichte ($s_{ik}(x) = \sum_j x_{ij} = x_{i+}$): Dieser Effekt berücksichtigt die Dichte[13] des Netzwerks bei der Schätzung der anderen Parameter. Er drückt die grundsätzliche Tendenz der Benutzer bzgl. des Auf- oder Abbaus von Verbindungen aus, weswegen er in allen Modellen einbezogen werden sollte. Ist der Koeffizient negativ, heißt dies im vorliegenden Fall, dass das Verfassen eines Reiseberichts für den Benutzer mit Kosten verbunden ist.

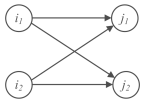

Abbildung 3: 4er-Kreis-Effekt

4er-Kreis ($s_{ik}(x) = \sum_{i_1,i_2,j_1,j_2} x_{i_1j_1} \cdot x_{i_1j_2} \cdot x_{i_2j_1} \cdot x_{i_2j_2}$): Drückt aus, wie stark sich Benutzer bei der Wahl der Reiseziele an anderen Benutzern orientieren, die in der Vergangenheit ähnliche Präferenzen bei Reisezielen gezeigt haben wie sie selbst (s. Abbildung 3). Ein positiver Koeffizient bedeutet, dass eine Beeinflussung durch die Entscheidungen Anderer stattfindet.

Popularität ($s_{ik}(x) = \sum_j x_{ij} \cdot x_{+j} = \sum_j x_{i_1j} \cdot \sum_{i_2} x_{i_2j}$): Beschreibt den Einfluss überdurchschnittlich vieler Reiseberichte für ein Reiseziel auf die zukünftige Wahl dieses Reiseziels. Ein positiver Koeffizient würde eine positive Rückkopplung anzeigen.

Aktivität ($s_{ik}(x) = x_{i+}^2$): Drückt den Einfluss überdurchschnittlich vieler Reiseberichte eines Benutzers auf die zukünftige Zahl seiner Reiseberichte aus. Ein positiver Koeffizient würde hier ebenfalls eine positive Rückkopplung anzeigen.

Dyadische Effekte, bezogen auf einen bestimmten Benutzer und ein bestimmtes Reiseziel. Formel für alle Effekte: $s_{ik}(x) = \sum_j x_{ij} \cdot (w_{ij} - \bar{w})$, $\bar{w}$: Mittelwert der w_{ij}

Bewertung: Beschreibt den Einfluss der durchschnittlichen Bewertung[14], die ein bestimmter Benutzer für ein bestimmtes Reiseziel vergeben hat. Ein positiver Koeffizient würde anzeigen, dass ein Benutzer ein Reiseziel, das er in einem Jahr überdurchschnittlich gut bewertet hat, im nächsten Jahr eher wieder besuchen wird.

Beiträge: Stellt den Einfluss der Anzahl der Diskussionsbeiträge dar, die ein bestimmter Benutzer zu einem bestimmten Reiseziel geschrieben hat. Im Falle eines positiven Koeffizienten würde ein Benutzer ein Reiseziel eher besuchen, wenn er überdurchschnittlich viele Diskussionsbeiträge zu ihm geschrieben hat.

Berichte: Drückt den Einfluss der Anzahl der Reiseberichte aus, die ein bestimmter Benutzer zu einem bestimmten Reiseziel geschrieben hat. Die Interpretation ist analog zum vorigen Effekt.

ego-Effekte ($s_{ik}(x) = v_i \cdot x_{i+} = v_i \cdot \sum_j x_{ij}$), Kovariateneffekte eines bestimmten Benutzers.

Alter: Beschreibt den Einfluss des Alters eines Benutzers auf seine Reiseaktivität. Ein positiver Koeffizient würde bedeuten, dass Benutzer, die älter als der Durchschnitt im jeweiligen Subnetzwerk sind, tendenziell eher reisen als Benutzer, die jünger sind; bei einem negativen Koeffizienten umgekehrt.

Geschlecht: Quantifiziert den Einfluss des Geschlechts eines Benutzers auf seine Reiseaktivität. Ein positiver Koeffizient würde anzeigen, dass Frauen eher reisen als Männer;

[12]Jeweils drei Kovariateneffekte der Benutzer, der Reiseziele und der Verbindung zwischen beiden beruhten auf der durchschnittlichen Bewertung in Reiseberichten, der Anzahl der Reiseberichte sowie der Anzahl der Diskussionsbeiträge. Um Korrelationen zu vermeiden, wurden Effekte, die auf der gleichen Größe beruhen, in diesem Schritt nicht kombiniert.

[13]Die Dichte eines Graphen ist das Verhältnis der Anzahl vorhandener Kanten in einem Graphen zur Anzahl maximal möglicher Kanten [24]. Die Anzahl der Kanten eines Graphen zum Zeitpunkt t ist $k_t = |E_t| = \sum_{ij} x_{ij}(t)$. Die maximal mögliche Anzahl Kanten ist $n \cdot (n-1)$ im gerichteten, nicht-partiten Fall und $n_1 \cdot n_2$ im bipartiten Fall. Die Dichte Δ ist dann $\Delta(t) = \frac{k_t}{n \cdot (n-1)}$ im nicht-partiten und $\Delta(t) = \frac{k_t}{n_1 \cdot n_2}$ im bipartiten Fall.

[14]Benutzer können im Reisebericht auch eine Bewertung des Freizeitangebots auf einer Skala von 1 (schlechteste Wertung) bis 5 (beste Wertung) vergeben.

Einflussfaktor	Subnetzwerk 1			Subnetzwerk 2		
	Modell 1-A	Modell 1-B	Modell 1-C	Modell 2-A	Modell 2-B	Modell 2-C
Strukturell						
Dichte	$-1,187\ (0,020)^{**}$	$-1,209\ (0,020)^{**}$	$-1,192\ (0,020)^{**}$	$-1,514\ (0,027)^{**}$	$-1,496\ (0,026)^{**}$	$-1,517\ (0,027)^{**}$
4er-Kreis				$0,022\ (0,002)^{**}$	$0,018\ (0,002)^{**}$	$0,021\ (0,002)^{**}$
Dyadisch						
Beiträge	$0,004\ (0,001)^{*}$	$0,004\ (0,001)^{*}$	$0,004\ (0,001)^{*}$			
Berichte	$0,033\ (0,007)^{**}$	$0,028\ (0,007)^{**}$	$0,032\ (0,007)^{**}$	$0,031\ (0,008)^{**}$	$0,030\ (0,007)^{**}$	$0,031\ (0,008)^{**}$
ego						
Alter	$-0,052\ (0,021)$	$-0,051\ (0,021)$	$-0,052\ (0,021)$			
Bewertung	$-0,099\ (0,048)$	$-0,107\ (0,047)$	$-0,098\ (0,048)$			
alter						
Bewertung	$-0,778\ (0,150)^{**}$	$-0,778\ (0,149)^{**}$	$-0,782\ (0,151)^{**}$	$-0,530\ (0,167)^{*}$	$-0,544\ (0,161)^{*}$	$-0,533\ (0,167)^{*}$
Beiträge	$0,207\ (0,011)^{**}$			$0,156\ (0,012)^{**}$		
Berichte		$0,308\ (0,015)^{**}$			$0,234\ (0,020)^{**}$	
Themen			$0,237\ (0,012)^{**}$			$0,179\ (0,015)^{**}$

Angabe: $\hat{\mu}$ ($\hat{\sigma}$); alle Schätzer signifikant zu $\alpha < 0,05$; *: $\alpha < 0,01$; **: $\alpha < 0,001$

Tabelle 3: Ergebnisse der finalen Modelle für die Subnetzwerke 1 und 2

Einflussfaktor	Subnetzwerk 3, Modelle A-C		Subnetzwerk 4, Modelle A-C		Subnetzwerk 5, Modelle A-C	
Strukturell						
Dichte	$[-2,172; -2,154]$	**	$[-1,667; -1,664]$	**	$[-1,723; -1,719]$	**
4er-Kreis	$[0,094; 0,101]$	**	$[0,221; 0,233]$	*	$[0,176; 0,181]$	*
Dyadisch						
Bewertung			$[0,420; 0,428]$		$[0,430; 0,434]$	
Berichte	$[0,142; 0,147]$	**	$[0,110; 0,114]$	*	$[0,171; 0,175]$	*
ego						
Bewertung					$[0,324; 0,325]$	
Beiträge	$[-0,080; -0,053]$	*				
Berichte	$0,102$	(nur 3-C)				
alter						
Beiträge	$0,230$	** (nur 3-A)	$0,139$	* (nur 4-A)	$0,103$	* (nur 5-A)
Berichte	$0,340$	** (nur 3-B)	$0,197$	* (nur 4-B)	$0,117$	* (nur 5-B)
Themen	$0,268$	** (nur 3-C)	$0,163$	* (nur 4-C)	$0,127$	* (nur 5-C)

Angabe: $[\hat{\mu}_{min}; \hat{\mu}_{max}]$; alle Schätzer signifikant zu $\alpha < 0,05$; *: $\alpha < 0,01$; **: $\alpha < 0,001$

Tabelle 4: Zusammenfassung der Ergebnisse der finalen Modelle für die Subnetzwerke 3, 4 und 5

bei einem negativen Koeffizienten umgekehrt.

Bewertung: Beschreibt den Einfluss des Mittelwerts aus allen Bewertungen für Reiseziele, die ein Benutzer für Reiseziele abgegeben hat. Ein positiver Koeffizient würde bedeuten, dass ein Benutzer, der im Mittel überdurchschnittlich gute Bewertungen vergeben hat, in Zukunft eher reisen wird als ein Benutzer, der im Mittel überdurchschnittlich schlechte Bewertungen vergeben hat.

Beiträge: Drückt den Einfluss der Anzahl der Diskussionsbeiträge (logarithmiert) aus, die ein Benutzer geschrieben hat. Bei einem positiven Koeffizienten wären Benutzer, die überdurchschnittlich viele Beiträge schreiben, tendenziell aktivere Reisende als Benutzer, die weniger Beiträge schreiben als der Durchschnitt.

Berichte: Stellt den Einfluss der Anzahl der Reiseberichte (logarithmiert) dar, die ein Benutzer geschrieben hat. Die Interpretation ist analog zum vorigen Effekt.

alter-Effekte ($s_{ik}(x) = \sum_j x_{ij} \cdot v_j$), Kovariateneffekte eines bestimmten Reiseziels.[15]

Bewertung: Beschreibt den Einfluss der durchschnittlichen Bewertung, die Freizeitangebote an diesem Reiseziel bekommen haben. Damit überdurchschnittlich gut bewertete Reiseziele eher gewählt werden, müsste der Koeffizient positiv sein.

Beiträge: Repräsentiert den Einfluss der Anzahl der Diskussionsbeiträge (logarithmiert), die von allen Benutzern zu einem Reiseziel geschrieben wurden. Diskussionsbeiträge lassen sich als Indikator dafür auffassen, wie stark sich die Benutzer der Reiseplattform mit einem Reiseziel beschäftigen, wie groß das Interesse an ihm ist, ohne dass tatsächlich Reisen stattgefunden haben müssen (wie es bei Reiseberichten angenommen wird). Bei einem positiven Koeffizienten wären überdurchschnittlich viele Beiträge zu einem Reiseziel ein Grund für die Benutzer, das Reiseziel zu wählen.

Berichte: Stellt den Einfluss der Anzahl der Reiseberichte (logarithmiert) dar, die von allen Benutzern zu Freizeitangeboten an einem Reiseziel geschrieben wurden. Die Interpretation des Koeffizienten ist analog zum vorigen Effekt.

Themen: Drückt den Einfluss der Anzahl der Diskussionsthemen (logarithmiert) aus, die zu einem Reiseziel eröffnet wurden. Die Anzahl der Diskussionsthemen lässt sich – ähnlich der Anzahl der Diskussionsbeiträge – als Indikator für das Interesse der Benutzer an dem Reiseziel interpretieren. Die Messgröße ist allerdings etwas unterschiedlich: Im Falle der Diskussionsthemen wird nicht zwischen langen und kurzen Diskussionen unterschieden, sondern eher die Spreizung des Interesses auf verschiedene Diskussionen bzw. Diskussionsthemen gemessen.

Beispielhaft sollen an dieser Stelle die Ergebnisse des Modells 2-A aus Tabelle 3 erläutert werden. Der geschätzte Parameter des Dichte-Effektes ist stark negativ ($\beta = -1,514$)

[15] Die Werte basieren auf dem gesamten Datensatz der Internetplattform, nicht nur auf den Daten des jeweiligen Subnetzwerks.

und statistisch höchst signifikant ($\alpha < 0,01$; $\sigma = 0,027$). Daraus lässt sich schließen, dass das Erstellen von Reiseberichten (bzw. damit auch Reisen) für die Akteure mit Kosten verbunden ist und nicht willkürlich erfolgt. Der positive ($\beta = 0,022$), statistisch höchst signifikante ($\alpha < 0,01$) 4er-Kreis-Effekt deutet darauf hin, dass Akteure, die in der Vergangenheit ähnliche Präferenzen bei ihren Reisezielen gezeigt haben, sich bei der Wahl neuer Reiseziele an den vergangenen Reisezielwahlentscheidungen der anderen Benutzer orientieren. Der positive ($\beta = 0,031$), statistisch höchst signifikante ($\alpha < 0,01$) dyadische Berichte-Effekt kann so interpretiert werden, dass ein Besucher in den folgenden Perioden ein Reiseziel dann besucht, wenn er in der Vorperiode besonders viele Reiseberichte zu diesem Reiseziel erstellt hat. Der negative ($\beta = -0,530$), statistisch schwach signifikante ($\alpha < 0,05$) Alter-Bewertung-Effekt impliziert, dass Reiseziele, die überdurchschnittlich gut bewertet werden, in folgenden Perioden eher nicht besucht werden. Der positive ($\beta = 0,156$) und statistisch höchst signifikante ($\alpha = 0,012$) alter-Beiträge-Effekt schließlich bedeutet, dass Reiseziele, zu denen in einer Vorperiode viele Diskussionsbeiträge im Forum erstellt wurden, in der nachfolgenden Periode bevorzugt besucht werden.
Die Modelle 2-B bzw. 2-C unterscheiden sich in der Auswahl der Effekte von 2-A nur dadurch, dass der alter-Beiträge-Effekt einmal durch den alter-Berichte-Effekt (2-B) und einmal durch den alter-Themen-Effekt (2-C) ersetzt wurden. Diese drücken die allgemeine Popularität eines Reiseziels anhand anderer Größen als der Anzahl der Diskussionsbeiträge aus. Die positiven ($\beta = 0,234$ bzw. $\beta = 0,179$) und statistisch höchst signifikanten ($\alpha < 0,01$) Effekte zeigen, dass die Anzahl Reiseberichte, die ein Reiseziel in der Vorperiode bekommen hat, sowie die Anzahl der Diskussionsthemen, die in der Vorperiode im Forum zu diesem Reiseziel verfasst wurden, einen positiven Einfluss auf die Wahl des Reiseziels in der Nachperiode haben. Für die alter-Effekte aus Beiträgen, Berichten und Themen mussten jeweils eigene Modelle gebildet, da die Korrelationen zwischen diesen Effekten ansonsten zu groß gewesen wären.

6. DISKUSSION

In den ausgewählten Subnetzwerken zeigten sich vor allem die Koeffizienten der Reisezieleffekte (alter-Effekte) als vergleichsweise stark und signifikant. Die Anzahl der Diskussionsbeiträge, Reiseberichte und Diskussionsthemen zu einem Reiseziel haben demnach einen positiven Einfluss auf die Wahrscheinlichkeit, dass das Reiseziel bzw. ein Freizeitangebot, das an diesem Reiseziel liegt, gewählt wird. Dies entspricht der intuitiven Erwartung, dass diese Größen die Popularität des Reiseziels ausdrücken.
Interessanterweise wirkten sich in den Subnetzwerken 1 und 2 eine überdurchschnittlich gute Bewertung des Reiseziels negativ (bzw. eine unterdurchschnittliche Bewertung positiv) auf die Wahrscheinlichkeit aus, was der negative Koeffizient für den Effekt alter–Bewertung anzeigt. Das widerspricht zunächst der intuitiven Erwartung. Wir haben hierfür zwei sich ergänzende Erklärungen. Zum einen ist der negative Einfluss überdurchschnittlich guter Bewertungen evtl. darauf zurückzuführen, dass solche Bewertungen von anderen Benutzern als gefälscht oder aus anderen Gründen als unrealistisch eingeschätzt wurden, sodass sie den Bericht ignorierten oder sich sogar deswegen gegen ein Reiseziel bzw. Freizeitangebot entschieden haben. Die Werte für eine durchschnittliche Bewertung lagen in etwa zwischen $3,4$ und $4,0$, mit einem Mittelwert von ca. $3,6$. Bei einer Bewertungsskala von 1 (schlechtester Wert) bis 5 (bester Wert) wurden also im Schnitt bereits gute Bewertungen vergeben.
Zum anderen fällt auf, dass dieser negative Einfluss bei den Subnetzwerken 3 bis 5 *nicht* vorhanden war, bei denen zusätzlich als Filterkriterium galt, dass die Benutzer zu den Top-1.000-Verfassern von Diskussionsbeiträgen gehören mussten. Wir vermuten, dass durch dieses zusätzliche Kriterium auch Benutzer ausgeschlossen wurden, die etwa selber Betreiber des Freizeitangebots sind, das sie auf der Internetplattform bewertet haben. Diese Benutzer hätten allen Grund, auf eine unterdurchschnittliche Bewertung ihres Freizeitangebots mit vermehrten Berichten zu reagieren, was jedoch dem Kalkül eines realen Reisenden widerspricht. Durch den Ausschluss der Betreiber in den Subnetzwerken 3 bis 5 wurden deren Entscheidungen in diesen Datensätzen nicht mitmodelliert und der negative Einfluss des alter-Bewertung-Effekts war nicht mehr in der Zielfunktion vorhanden.

In den finalen Modellen aller Subnetzwerke war der dyadische Effekt der Anzahl der Berichte mit schwachem bis mittelmäßig starkem Einfluss enthalten. Der dyadische Effekt der durchschnittlichen Bewertung zeigte immerhin noch in zwei der fünf Subnetzwerke einen signifikanten, stark positiven Einfluss, der der Diskussionsbeiträge war nur in einem Netzwerk signifikant und hatte nur einen sehr schwachen, positiven Einfluss.
Die dyadischen Effekte zeigen an, dass Benutzer stärker dazu neigen, bereits im Vorjahr besuchte Reiseziele erneut zu besuchen, wenn sie zu ihnen Reiseberichte oder Diskussionsbeiträge schreiben oder sie in Reiseberichten gut bewerten. Dies entspricht der Erwartung, dass zufriedene (gute Bewertung) oder weiterhin interessierte (viele Diskussionsbeiträge) Benutzer ein Reiseziel erneut besuchen.

Die Effekte aus Eigenschaften der Benutzer hatten in den Subnetzwerken insgesamt wenig Einfluss. In keinem finalen Modell hatte das Geschlecht der Benutzer einen Einfluss, das Alter nur in einem Subnetzwerke und sein Einfluss war dort auch nur sehr schwach. Ebenso hatten die Anzahl der Diskussionsbeiträge und der Reiseberichte, die ein Benutzer verfasst hat, nur in den finalen Modellen eines Subnetzwerks einen schwachen Einfluss. Der Einfluss der durchschnittlichen Bewertung, die ein Benutzer vergeben hat, war zwar in zwei Subnetzwerken in den finalen Modellen enthalten, allerdings einmal mit positivem, einmal mit negativem Vorzeichen, wobei die Korrelation dieses Effekts mit anderen Effekten in den finalen Modellen sehr gering war (Korrelationskoeffizient absolut kleiner $0,13$). Die Konsistenz dieses Einflusses ist daher fraglich.

In der Gruppe der strukturellen Effekte war der Effekt der Dichte in allen Modellen enthalten und hatte einen starken negativen Einfluss. Dies untermauert die Vermutung, dass das Verfassen von Reiseberichten für den Benutzer mit Kosten verbunden ist, da das Reisen selbst Kosten verursacht. Von den übrigen drei Effekten war nur der 4er-Kreis-Effekt in den finalen Modellen von vier Subnetzwerken enthalten, dort mit schwachem bis mittelmäßig starkem Einfluss. Er zeigt an, dass Benutzer dazu neigen, Reiseziele zu besuchen, die andere Benutzer mit ähnlichen Reisezielpräferenzen bereits besucht haben.

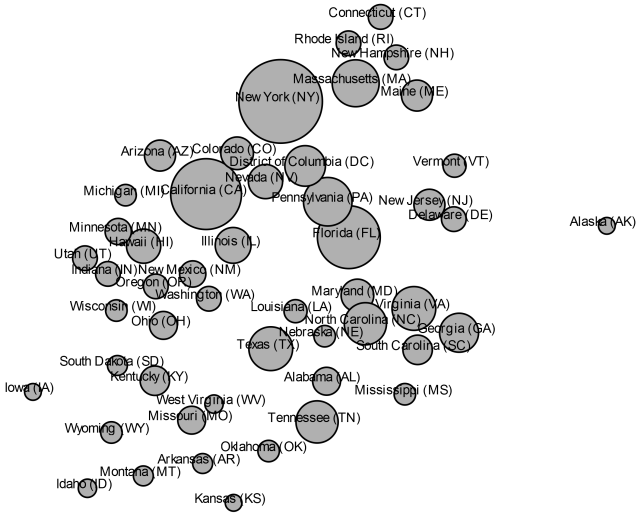

Abbildung 4: „Nähe“ der Reiseziele zueinander anhand des Reiseverhaltens der Benutzer im Jahr 2009 für das Subnetzwerk 2

Die Ergebnisse zeigen, dass der gewählte Ansatz vielversprechend ist und Potenzial für weitere Untersuchungen bietet. Auf der Reiseplattform beeinflussen – neben den grundsätzlichen Kosten einer Reise – vor allem Eigenschaften der Reiseziele, aber auch die Beziehung eines Benutzers zu einem Reiseziel (dyadische Effekte) und die Reisen anderer Benutzer mit ähnlichen Präferenzen die Reiseentscheidungen.

Der Einfluss einzelner unabhängiger Variablen auf die Chance (Odds[16]), dass ein Benutzer als nächstes einen Reisebericht zu einem bestimmten Reiseziel schreibt, wird in der logistischen Regression durch den Effektkoeffizienten e^{β_k} beschrieben. Bei Erhöhung der unabhängigen Variable um $+1$, erhöht sich der Odds um den Faktor e^{β}.
Am Beispiel des alter-Beiträge-Effekts, dessen Koeffizienten in der Größenordnung von ca. $\beta = 0,2$ lagen, bedeutet das, dass sich die Chance um den Faktor $e^{\beta} = e^{0,2} \approx 1,2214$ erhöht, also um ca. $22,14$ Prozent, wenn die unabhängige Variable um $+1$ größer wird. Dabei ist zu beachten, dass die Werte dieses Effekts vor der Analyse in RSiena logarithmiert wurden (Basis e), so dass sich nicht die Anzahl der Diskussionsbeiträge um $+1$ erhöhen muss, sondern deren Logarithmus. Die Erhöhung des Logarithmus' um $+1$ entspricht der Multiplikation des ursprünglichen Werts mit e^1. Angenommen, ein Reiseziel hat bisher 100 Diskussionsbeiträge erhalten und der Koeffizient für alter–Beiträge sei $\beta = 0,2$, dann müsste dieses Reiseziel $100 \cdot (e^1 - 1) \approx 172$ Beiträge mehr bekommen, damit sich seine Chance (Odds), gegenüber anderen Reisezielen gewählt zu werden, um $22,14$ Prozent erhöht. Hätte es bisher 100.000 Diskussionsbeiträge erhalten, wäre die absolute Zahl zusätzlich notwendiger Beiträge entsprechend größer.
Für die anderen Effekte gilt diese Rechnung analog, wobei bei Effekten, deren Daten nicht logarithmiert wurden, die Erhöhung der unabhängigen Variable direkt der Erhöhung der entsprechenden Größe im Datensatz entspricht.

[16]Verhältnis der Wahrscheinlichkeit, dass ein Ereignis eintritt, zur Gegenwahrscheinlichkeit, dass es nicht eintritt.

Abbildung 4 visualisiert exemplarisch die „Nähe“ der Reiseziele zueinander, so wie sie sich aus dem Reiseverhalten bzw. den entsprechenden Reiseberichten der 100 Benutzer des Subnetzwerkes 2 im Jahr 2009 ergibt. Je näher zwei Reiseziele in der Abbildung nebeneinander liegen, desto öfter wurden beide Reiseziele von denselben Benutzern bewertet. Diese Darstellung spiegelt somit auch den 4er-Kreis-Effekt wider.
Die Größe des zugehörigen Kreises zu einem Reiseziel illustriert die Gesamtanzahl der Reiseberichte, die das jeweilige Reiseziel im Jahr 2009 von diesen 100 Benutzern erhalten hat. Die meisten Bewertungen im Jahr 2009 erhielt die Stadt New York. Auch wenn New York und Kalifornien geographisch weit voneinander entfernt liegen, sind sie in der Abbildung jedoch sehr nahe zueinander angeordnet, da dieselben Benutzer der online Plattform sowohl Reiseziele in New York als auch in Kalifornien bewertet haben. Die 100 Benutzer, die diese Reiseberichte erstellt haben, sowie die Kanten zwischen Nutzern und Reisezielen wurden aus der Abbildung ausgeblendet, um deren Lesbarkeit zu erhöhen.

6.1 Theoretische und praktische Relevanz

Aus wissenschaftlicher Sicht ist die Untersuchung in zweierlei Hinsicht relevant. Erstens ist der gewählte Ansatz die erste Anwendung einer inferenzstatistischen Methode auf die Dynamik eines bipartiten Netzwerks zur Untersuchung des Reiseverhaltens von Akteuren. Dabei konnte gezeigt werden, dass sich diese Fragestellung mit Hilfe der Klasse der akteursbasierten Modelle grundsätzlich untersuchen lässt.
Zweitens bietet der Ansatz die Möglichkeit, die inhaltlichen Ergebnisse anderer Studien hinsichtlich der Einflussfaktoren für Reiseentscheidungen zu überprüfen und gegebenenfalls qualitative Ergebnisse zu quantifizieren.
Der Ansatz lässt sich darüber hinaus auch auf andere Bewertungsportale im Internet übertragen, um Entscheidungen in Bezug auf Produkte oder Dienstleistungen anderer Domänen zu untersuchen. So kann beispielsweise der 4er-Kreis-Effekt Einflüsse durch Ähnlichkeiten in den Präferenzen verschiedener Benutzer / Käufer bzgl. des gleichen Buches, des gleichen Films oder der gleichen CD aufdecken. Dyadische Effekte eignen sich besonders, um Einflussfaktoren für Wiederholungskäufe z. B. bei Lebensmitteln oder Haushaltswaren zu identifizieren. Mittels der alter-Effekte lässt sich die allgemeine Popularität (möglicherweise in verschiedenen Größen gemessenen) von Produkten für eine Gruppe von Käufern abbilden und ihr Einfluss analysieren. Das kann beispielsweise bei solchen Elektronikartikeln eine wichtige Größe sein, bei denen weniger persönliche Präferenzen oder bisherige Erfahrung mit dem Produkt eine Rolle spielen, als viel mehr, dass das Produkt sich bei vielen Käufern bewährt hat.

In der Praxis kann der vorgestellte Ansatz vor allem für die Reise- und Tourismusbranche interessant sein. Die Weiterentwicklung des Ansatzes zu einer Prognosemethode für Reiseströme kann Unternehmen wie Fluggesellschaften helfen, die zur Planung ihrer zukünftigen Allokation von Personal, Flugzeugen und Zielflughäfen auf möglichst genaue Vorhersagen auf diesem Gebiet angewiesen sind.
Für Unternehmen, die selber als Freizeitangebote auf der Reiseplattform vertreten sind, besteht die Möglichkeit, ihre aktuelle Attraktivität auf der Plattform quantitativ einzuschätzen.

6.2 Kritische Würdigung und zukünftige Forschung

Das Ziel, einen Ansatz zu entwickeln, um Einflussfaktoren für Reiseentscheidungen von Benutzern einer online Reiseplattform zu identifizieren, wurde erreicht. Anhand von fünf Beispielnetzwerken wurden Modelle entwickelt und geschätzt, die jeweils den Einfluss verschiedener Effekte auf die Reiseentscheidung quantifizieren. Dabei konnten in jedem Modell mehrere signifikante Effekte isoliert werden. Die identifizierten Einflussfaktoren lassen sich jedoch aufgrund der (bewussten) starken Eingrenzung auf sehr aktive Benutzer nicht auf alle Benutzer der Internetplattform übertragen.

Wie jede empirische Studie, ist auch diese an einige zentrale Annahmen geknüpft, die ihre Aussagekraft unter Umständen einschränken könnten. Die zentralste Annahme bei der Auswertung von Reiseberichten aus dem Internet ist, dass ein Reisebericht eine reale Reise des Verfassers des Berichts zu dem bewerteten Freizeitangebot und damit zu dem entsprechenden Reiseziel widerspiegelt. Es besteht jedoch grundsätzlich die Möglichkeit, dass z. B. der Betreiber eines Freizeitangebots selber Reiseberichte darüber schreibt. Abgesehen davon, dass diese Berichte inhaltlich fragwürdig sind, da der Betreiber sein Angebot wohl immer gut bewerten wird, liegen ihnen keine realen Reisen zugrunde. Dadurch verzerren sie die Beobachtung des Reiseverhaltens anhand der Internetplattform. In unserer Untersuchung haben wir versucht, gefälschte Berichte herauszufiltern, indem wir aktive Benutzer fokussiert haben, die Berichte zu verschiedenen Reisezielen verfasst haben, und in drei der fünf Subnetzwerke zusätzlich verlangt haben, dass die Benutzer häufig Diskussionsbeiträge in Foren schreiben. Dennoch sollten weitere Forschungen dieses Problem untersuchen.

Außerdem sollten insbesondere die Gründe für den negativen Einfluss einer überdurchschnittlich guten Bewertung eines Reiseziels auf seine Wahlwahrscheinlichkeit näher untersucht werden. Wie dargestellt nehmen wir an, dass dieser Einfluss dadurch zustande kommt, dass Benutzer der Internetplattform außerordentlich gute Reiseberichte für ein Reiseziel als gefälscht oder naiv ansehen, und / oder, dass Betreiber von Freizeitangeboten auf schlechte Bewertungen mit (positiven) Berichten reagieren. Um diese Annahmen zu überprüfen, erscheint es sinnvoll, bei Reiseberichten den Zusammenhang zwischen überdurchschnittlicher Bewertung, Bewertung des Berichts durch andere Benutzer und Text (Länge, Semantik) zu untersuchen.

Daneben können die in diesem Artikel angewandten Modelle um zusätzliche Effekte erweitert werden, da die gesammelten Daten mehr Informationen beinhalten, als bisher verwendet wurden. So wurden etwa die Informationen über die Herkunft der Benutzer in dieser Untersuchung nicht genutzt, da das entsprechende Feld im Benutzerprofil nicht standardisiert ist, sondern Freitext zulässt. Eine nachträgliche Standardisierung der Angaben böte die Möglichkeit, die Reiseentscheidungen in Herkunft-Ziel-Reiseströme einzuordnen – eine Betrachtung, die für die Vorhersage der Nachfrage in der Reisebranche von großer Bedeutung ist.

Um den Ansatz dieses Artikels zu einer Prognosemethode für Reiseströme bzw. -trends zu erweitern, ist es außerdem notwendig, auf Basis der finalen Modelle zunächst Vorhersagen über die Entwicklungen von Reiseenscheidungen innerhalb des Datensatzes zu entwickeln und diese dann zu überprüfen. Die Ergebnisse sollten dann mit Realdaten zu Reiseentscheidungen und Reiseströmen verglichen und die Methode weiter automatisiert werden. Wir hoffen, dass diese Studie als Basis für derartige Nachfolgestudien dient.

7. REFERENCES

[1] M. Ben-Akiva and S. Lerman. *Discrete choice analysis: theory and application to travel demand.* MIT Press, 1985.

[2] K. de Valck. *Virtual Communities of Consumption: Networks of Consumer Knowledge and Companionship.* PhD thesis, Erasmus University Rotterdam, 2005.

[3] R. Diestel. *Graphentheorie.* Springer-Verlag, Heidelberg, 3. edition, 2006.

[4] O. Frank and D. Strauss. Markov graphs. *Journal of the American Statistical Association*, 81(395):832 – 842, 1986.

[5] E. García-Barriocanal, M.-A. Sicilia, and N. Korfiatis. 117 exploring hotel service quality experience indicators in user-generated content: A case using tripadvisor data. In *MCIS 2010 Proceedings*, 2010. Paper 33.

[6] P. Holland and S. Leinhardt. A dynamic model for social networks. *Journal of Mathematical Sociology*, 5(1):5 – 20, 1977.

[7] P. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373):33 – 50, 1981.

[8] P. Jaccard. Contributions au problème de l'immigration post-glaciaire de la flore alpine. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547 – 579, 1900.

[9] M. Latapy, C. Magnien, and N. D. Vecchio. Basic notions for the analysis of large two-mode networks. *Social Networks*, 30(1):31 – 48, 2008.

[10] L. Mendes-Filho and F. B. Tan. User-generated content and consumer empowerment in the travel industry: A uses & gratifications and dual-process conceptualization. In *PACIS 2009 Proceedings*, 2009. Paper 28.

[11] J. Miguéns, R. Baggio, and C. Costa. Social media and tourism destinations: Tripadvisor case study. In *International Association for the Scientific Knowledge – Advances in Tourism Research*, 2008.

[12] M. O'Mahony, P. Cunningham, and B. Smyth. An assessment of machine learning techniques for review recommendation. In *20th Irish Conference on Artificial Intelligence and Cognitive Science*, 2009.

[13] M. O'Mahony and B. Smyth. A classification-based review recommender. *Knowledge-Based Systems*, 23(4):323 – 329, 2010.

[14] P. Pattison and S. Wasserman. Logit models and logistic regressions for social networks: Ii. multivariate relations. *British Journal of Mathematical and Statistical Psychology*, 52(2):169 – 193, 1999.

[15] J. Putzke, K. Fischbach, D. Schoder, and P. Gloor. The evolution of interaction networks in massively multiplayer online games. *Journal of the Association*

for Information Systems, 11:69 – 94, 2010.

[16] U. Rabanser and F. Ricci. Recommender systems: Do they have a viable business model in e-tourism? In A. J. Frew, editor, *Information and Communication Technologies in Tourism 2005*, pages 160 – 171. Springer Vienna, 2005.

[17] F. Ricci and R. Wietsma. Product reviews in travel decision making. In M. Hitz, M. Sigala, and J. Murphy, editors, *Information and Communication Technologies in Tourism 2006*, pages 296 – 307. Springer Vienna, 2006.

[18] T. Snijders. Stochastic actor-oriented dynamic network analysis. *Journal of Mathematical Sociology*, 21:149 – 172, 1996.

[19] T. Snijders. The statistical evaluation of social network dynamics. *Sociological Methodology*, 31(1):361 – 395, 2001.

[20] T. Snijders. Models for longitudinal network data. In P. Carrington, J. Scott, and S. Wasserman, editors, *Models and Methods in Social Network Analysis*, pages 215 – 247. Cambrige University Press, New York, 2005.

[21] T. Snijders, G. van de Bunt, and C. Steglich. Introduction to stochastic actor-based models for network dynamics. *Social Networks*, 32(1):44 – 60, 2009.

[22] H. Song and G. Li. Tourism demand modelling and forecasting–a review of recent research. *Tourism Management*, 29(2):203 – 220, 2008.

[23] C. Steglich and A. Knecht. Die statistische Analyse dynamischer Netzwerkdaten. In C. Stegbauer and R. Häußling, editors, *Handbuch der Netzwerkforschung*. Verlag für Sozialwissenschaften, Wiesbaden, 2010.

[24] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge, 1994.

[25] S. Wasserman and P. Pattison. Logit models and logistic regressions for social networks: I. an introduction to markov graphs and p*. *Psychometrika*, 61(3):401 – 425, 1996.

[26] S. Witt and C. Witt. Forecasting tourism demand: A review of empirical research. *International Journal of Forecasting*, 11(3):447 – 475, 1995.