Netzwerk-Perspektive

Über die Erhebung und Analyse sozialer Online-Kommunikation

Pascal Jürgens, IFP Mainz

Takeaways

- Warum Netzwerke?
- Online-Kommunikation und Verhalten
- Neue Datenquellen - Blogs, SNS, Foren, IM…
- Regulierungs-Powerplay - Datensilos
- Fremde Spielregeln - APIs
- Sampling-Sorgen
- Datenhaushalt
- Erstellung eines Netzwerks
- Visualisierung
- Deskriptive Analysen
- Inferenz (ERGMs)
- Zusammenfassung / Fragen

0

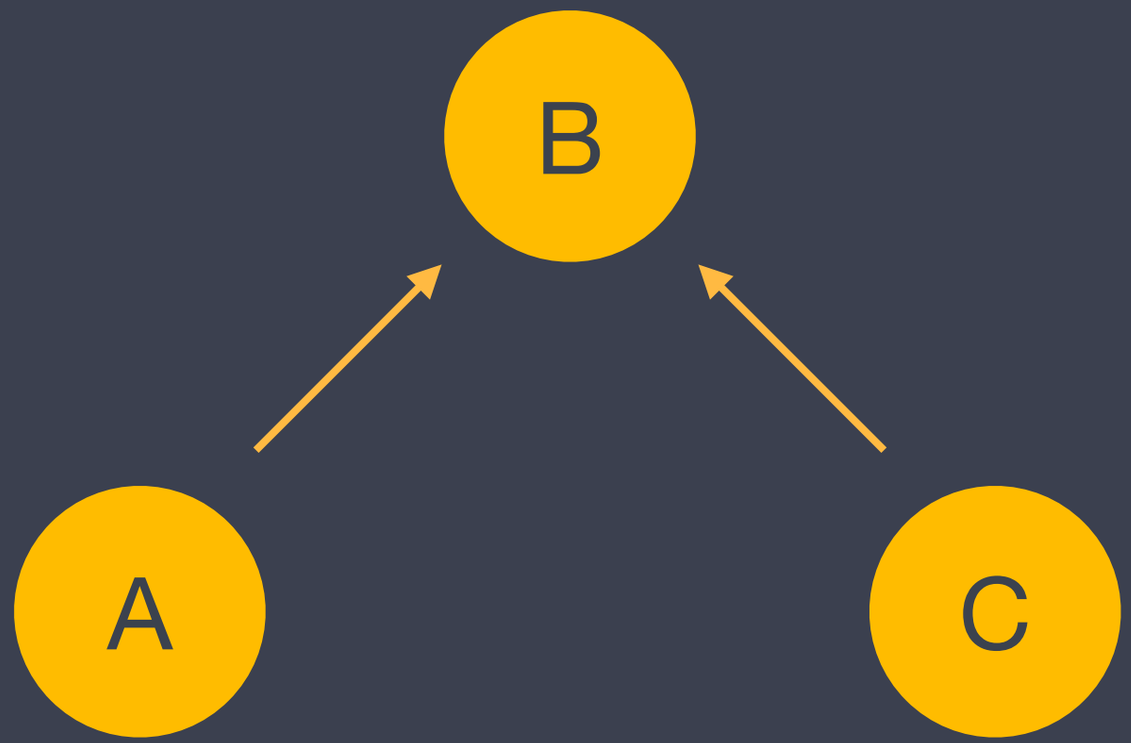
Warum Netzwerke?

http://www.visualcomplexity.com/vc/

Warum Netzwerke?

1. Sehen / Zeigen von sozialen Kommunikationsstrukturen (Wer ist wichtig? Wer steht am Rand?)
2. Gruppenbildung (Communities)
3. Beschreiben ganzer Kommunikationsnetze (Wie schnell/gut wird Information weitergeleitet? Wie dicht ist das Netzwerk? Wie ist Prominenz/Sichtbarkeit verteilt?
4. Konfirmative Analyse (Hypothesentests) über die Entstehung der Netzwerkstruktur

Beispiel



1

Kommunikation / Verhalten

To date, the methods employed in Internet research have served to critique the persistent idea of the Internet as a virtual realm apart.

Rogers, R. (2009). The End of the Virtual. Digital Methods (S. 5). Amsterdam University Press.

I would like to suggest inaugurating a new era in Internet research, which no longer concerns itself with the divide between the real and the virtual. It concerns a shift in the kinds of questions put to the study of the Internet. The Internet is employed as a site of research for far more than just online culture. The issue no longer is how much of society and culture is online, but rather how to diagnose cultural change and societal conditions using the Internet. The conceptual point of departure for the research program is the recognition that the Internet is not only an object of study, but also a source.

Rogers, R. (2009). The End of the Virtual. Digital Methods (S. 8). Amsterdam University Press.

Kommunikation vs. Verhalten

- Vormals: Massenkommunikation
- Inzwischen: Online-Kommunikation ist oft Individualkommunikation aber öffentlich und manifest — **und**
- besitzt Handlungsqualitäten (Beziehungspflege, gesellschaftliche Teilhabe, Handel …)


Stuttgart 21
www.kopfbahnhof-21.de
Demokratie21

Folgen

- Neue Untersuchungsgegenstände
- Neue Datenquellen
- Neue Methoden?

2

Neue Datenquellen

Neue Qualitäten

- Manifeste Online-Kommunikation oder digitale Spuren (Rogers) besitzen neue Charakteristiken:
- Non-reaktiv
- (idR.) temporal lokalisierbar
- Maschinell erfassbar
- Semi-strukturiert
- **Relational ➜ Netzwerkstruktur!**

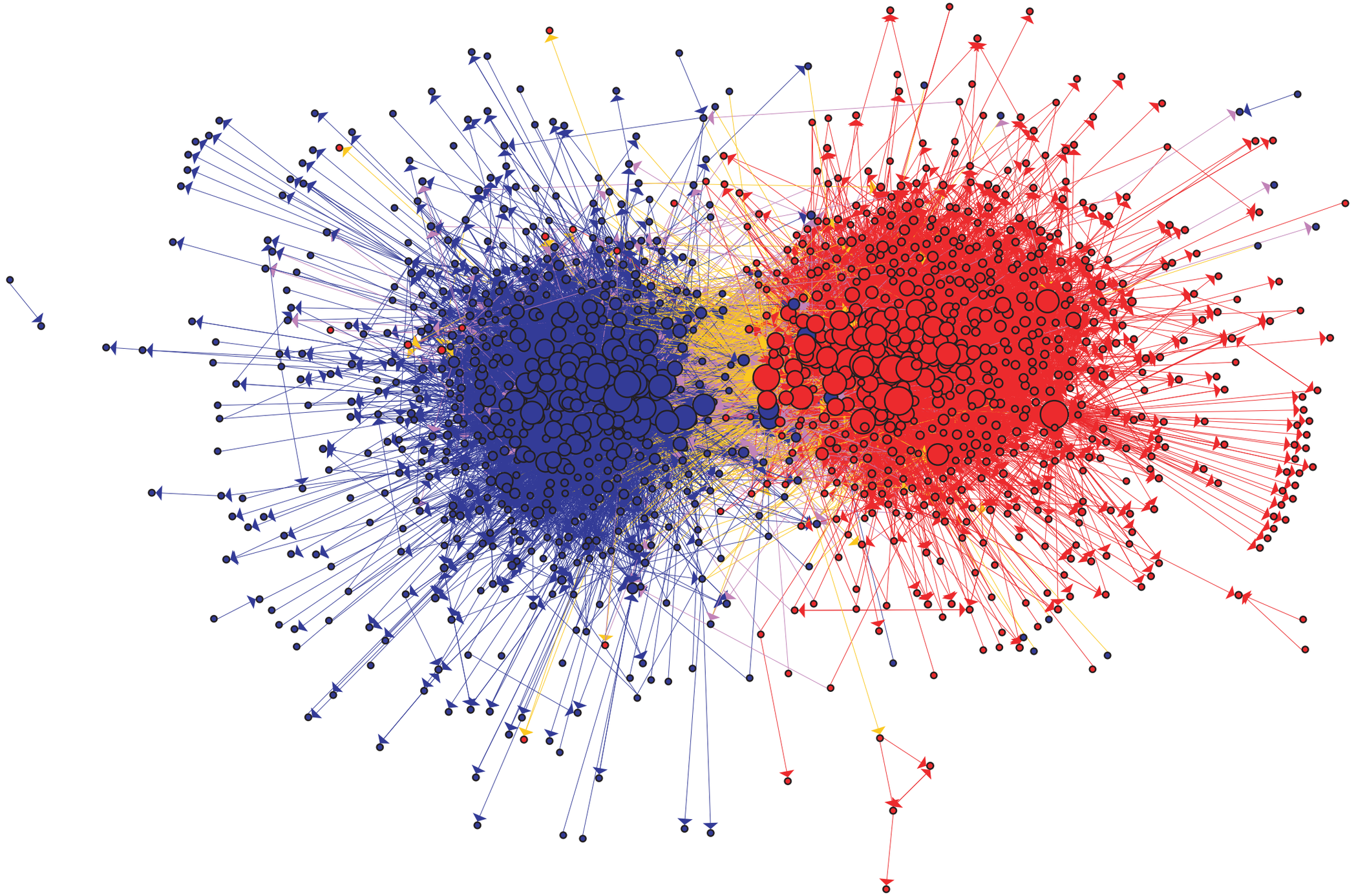
What makes a Network

These datasets have given birth to innovative and substantively diverse publications, all premised on the "anticategorical imperative" (Emirbayer and Goodwin, 1994, p. 1414) which privileges relations over categorical attributes in the explanation of social behavior.

Lewis, K., Kaufman, J., Gonzalez, M., Wimmer, A., & Christakis, N. (2008). Tastes, ties, and time: A new social network dataset using Facebook. com. Social Networks, 30(4), 330–342. Elsevier.

Beispiele für relationale Daten

- Links (Hyperlinks)
- Weblogs (Zitationsnetzwerke)
- Zitate / Reaktionen in Foren
- “*Freundschaften*” in SNS (Social Network Sites)
- Email-Austausch
- …



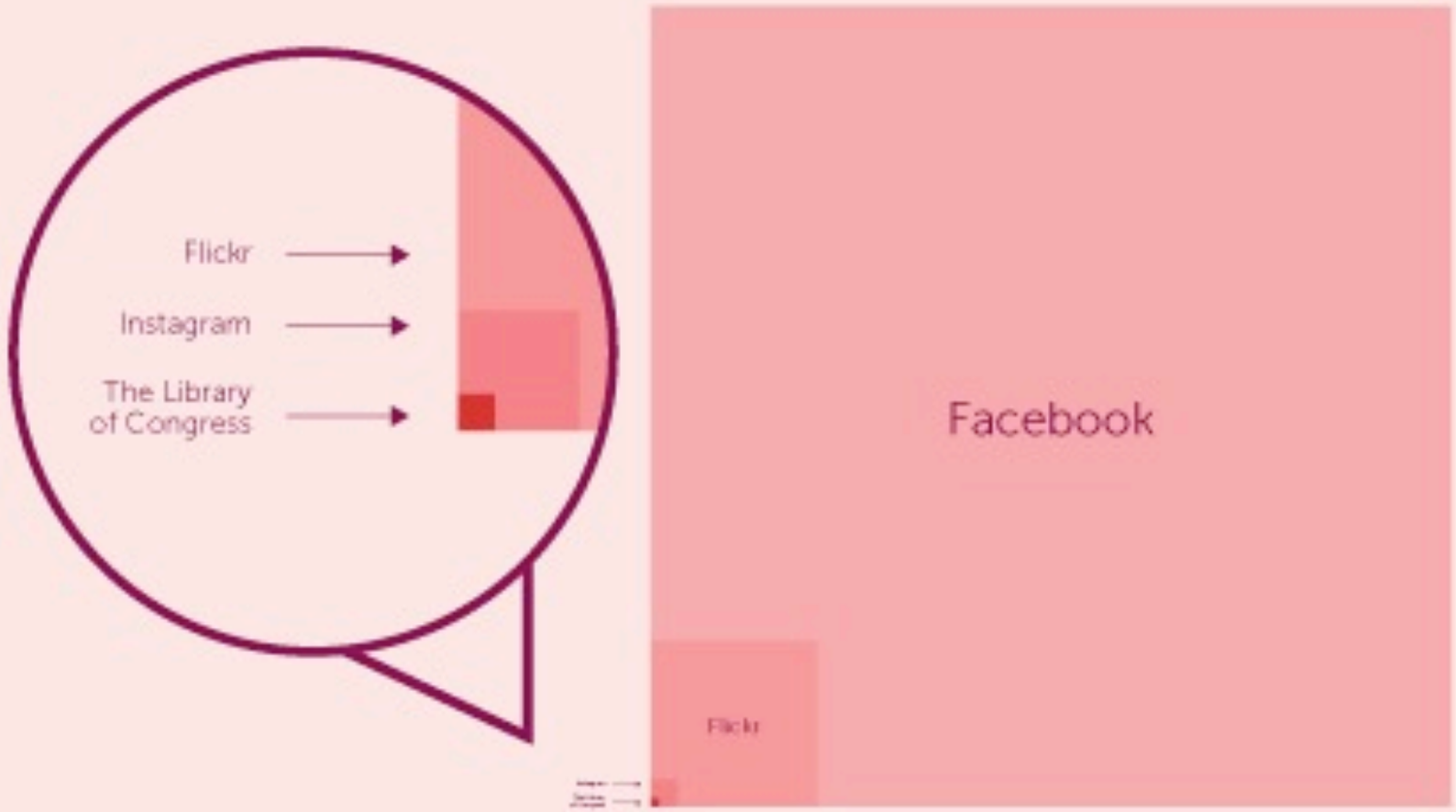
Adamic, L. A., & Glance, N. (2005). The political blogosphere and the 2004 US election: divided they blog. Proceedings of the 3rd international workshop on Link discovery, 36–43. ACM.

3 Regulierungs-Powerplay und Datensilos

But computational social science is occurring—in Internet companies such as Google and Yahoo, and in government agencies such as the U.S. National Security Agency. Computational social science could become the exclusive domain of private companies and government agencies. Alternatively, there might emerge a privileged set of academic researchers presiding over private data from which they produce papers that cannot be critiqued or replicated. Neither scenario will serve the long-term public interest of accumulating, verifying, and disseminating knowledge.

Lazer, D. et al. (2009). Computational Social Science.

THE WORLD'S LARGEST PHOTO LIBRARIES



Herausforderung für die Zukunft

- Generierung, Erfassung und Auswertung von Nutzerdaten sind ein Wachstumsmarkt
- Daten und Erkenntnisse sind für Firmen direkt äquivalent zu Geld (Werbung, Recommendations)
- Zunehmend restriktivere Handhabung, auch gegenüber der Wissenschaft (AOL-Datenset etc.)
- Notwendig: Eigene Kompetenzen, eigene Datenerhebung

4

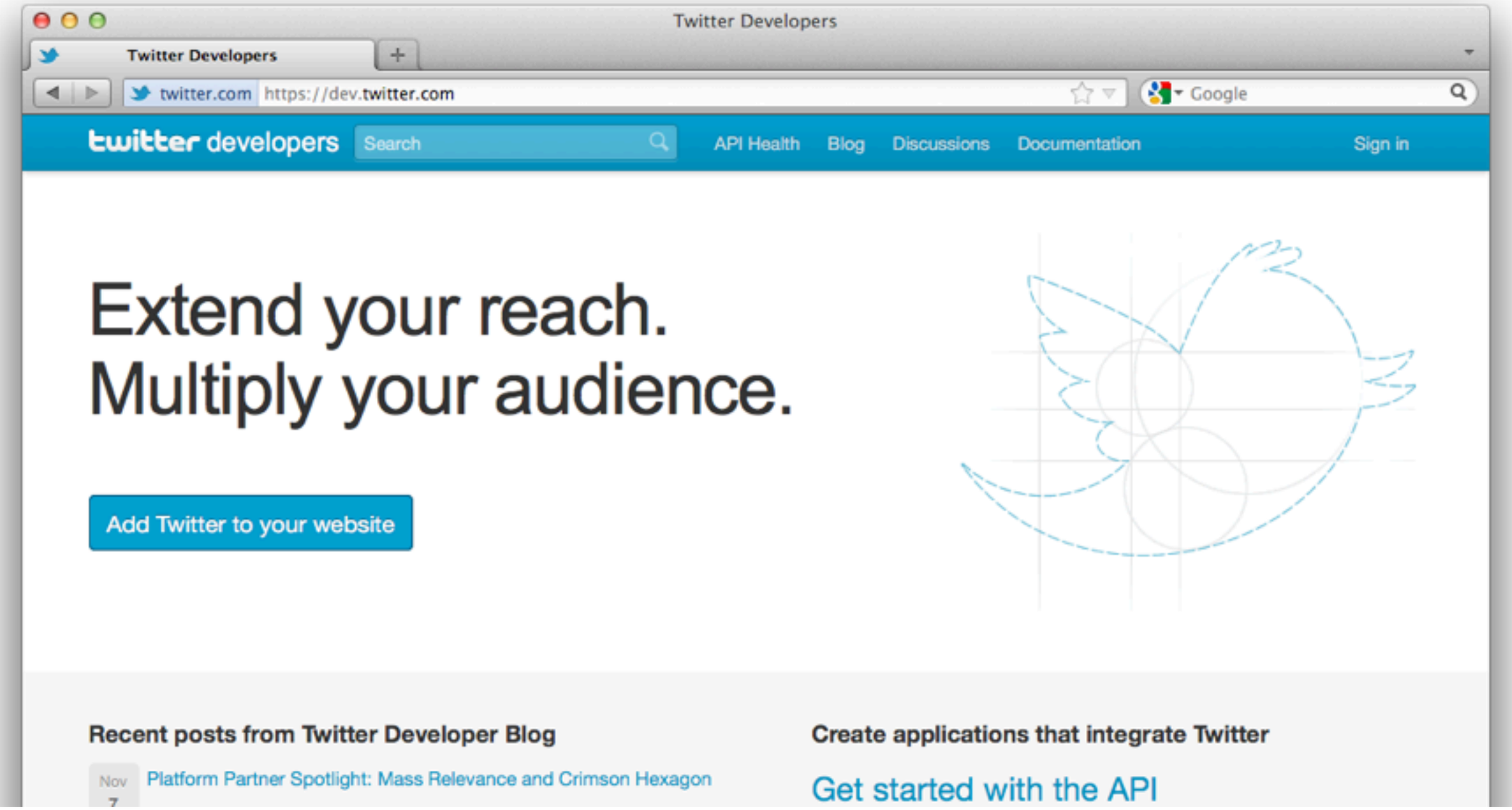
Fremde Spielregeln – APIs

Warum APIs?

- Inhalte sind reaktiv / Änderungen unterworfen (maschinelle Erhebung ist deterministisch)
- Copy & Paste / Screenshots sind ab 3-4 Seiten zu viel Arbeit
- Abgespeicherte Seiten sind nicht strukturierte Daten ➜ Zusatzaufwand bei der Analyse

APIs (Application Programming Interface)

- API = programmatischer Zugriff auf Ressourcen
- Fast immer mit Registrierung / Nutzungsbeschränkung
- Wer sich nicht an die Regeln hält wird gesperrt
- Liefern klar definierte Datensets
- Müssen über eigene Programme abgefragt werden
  (☞ Python lernen oder Programmierer finden)
- Sind für Programmierer konzipiert, nicht für Wissenschaftler



Twitter: API-Dokumentation

Beispiel: Twitter-API

- Jeder Twitter-Nutzer kann bis zu 350 Anfragen pro Stunde senden
- Abrufen von Tweets, Informationen zu Nutzern, Freundes- und Follower-Listen
- Privilegierte Entwickler erhalten höhere Kontingente (früher umsonst, heute idR. kostenpflichtig)
- Zugriff auf echtes Zufalls-Sample an Tweets (!)

Beispiel: Daten in einem Tweet

https://dev.twitter.com/docs/api/1/get/statuses/home_timeline

https://dev.twitter.com/docs/api/1/get/followers/ids

Twitter-API Demo

Facebook: API-Dokumentation

Beispiel: Facebook-API

- Jeder Facebook-Entwickler (Anmeldung) kann eine unbezifferte Anzahl an Anfragen pro Stunde senden
- Zugänglich sind grundsätzlich alle öffentlichen Daten (Profile, Wall-Posts, Fotos etc.)
- Zugriff auf private Inhalte ist nach Zustimmung von Nutzern möglich

Facebook-API Demo

Which types of research would be ‘scooped’ by Facebook’s flipping a switch? Facebook serves as one notable example of the sudden reconfiguration of a research object, which is common to the medium.

Rogers, R. (2009). The End of the Virtual. Digital Methods (S. 10). Amsterdam University Press.

5

Sampling-Sorgen

Problem: Stichproben

- Unbekannte Grundgesamtheiten im Internet
- Fehlende zufällige Zugriffsmethoden
- Eingeschränkte Zugriffsrechte
- Ausnahmen: Blogs (Snowball ➜ Grundgesamtheit ➜ Stichprobe)
- Ausnahmen: Twitter (Sample-Stream)
- Neuere Methoden: Gjoka, M., Kurant, M., Butts, C. T., & Markopoulou, A. (2010). Walking in Facebook: A case study of unbiased sampling of OSNs. INFOCOM, 2010 Proceedings IEEE, 1–9. Ieee.

6 Datenhaushalt

Nebenschauplatz: Datenmanagement

- Nicht trivial:
- Daten speichern (Datenbank?) ➜ Speicherplatz
- Daten aufbereiten ➜ Zeitaufwand
- Daten in Netzwerk umwandeln

Nebenschauplatz: Privatsphäre

- Faustregel:
- **Niemals Datensätze veröffentlichen**
- Grund: AOL-Datensatz, Netflix-Datensatz, 3T-Datensatz
- Zimmer, M. (2010). “But the data is already public”: on the ethics of research in Facebook. Ethics and Information Technology, 12(4), 313–325. doi:10.1007/s10676-010-9227-5

7

Erstellung eines Netzwerks

Netzwerkgenerierung in drei Schritten

- 1 - Was ist ein Link?
  (Twitter: @-Nachricht, Following? Facebook: Freundschaft? Reziprokes Kommentieren?)
- Vorsicht bei heterogenen Inhalten! (geht es um die Verbreitung eines spezifischen Inhaltes oder vieler?)
- 2 - Wie lange ist der Aggregationszeitraum?
  (logische Größe: 1 Tag? Kampagnendauer?)
- 3 - Sonstige Parameter (gerichtete Links? Gewichte?)

Pseudo-Code

- Betrachte jeden empirischen Datenpunkt, der eine Relation etabliert
- Füge die beschriebene Relation zwischen Person A und B dem Netzwerk hinzu
  (Falls A oder B nicht existieren, füge sie ebenfalls hinzu)
- Ergänze beschreibende Merkmale der Personen
  (z.B. Geschlecht, Alter aber auch Parteiaffinität etc.)

Beispiel aus meinem Code

```
for tweet in tweets.find({'to': 'username '}):
    edges[tweet['from'], tweet['to']] += 1
```

8
Visualisierungen

Achtung!

- Netzwerk-Visualisierungen sind grundsätlich nicht konfirmativ einsetzbar!
- Funktionsweise: Physikalische Modelle - Verbindungen sind Federn zwischen Knoten
- Ziel sollte sein: Übersichtliche Darstellung der Struktur, kohärente Subgruppen
- Nützlich: Farben für Knoteneigenschaften, Größe für Wichtigkeit / Prominenz

Visualisierungs-Demo

9

Deskriptive Analysen

Kennwerte: Netzwerklevel

- Größe des Netzwerks
- Dichte des Netzwerks (Grad der Verknüpfung)
- Duchmesser (➜ Geschwindigkeit der Informationsverbreitung)
- Clustering-Koeffizient C_i ➜ Anteil aller denkbaren Dreiecks-Verbindungen (Triple), die geschlossen sind (alle drei Knoten miteinander verbunden)

Kennwerte: Akteurs-/ Knotenebene

- Großer Forschungsfokus: Maße für Einfluss
- “Grad” (*Degree*) — Anzahl der Verbindungen (auch separat für ein- bzw. ausgehend: *Indegree vs Outdegree*)
- Zentralität (*Centrality*) — Wichtigkeit für Informationsfluss (sozusagen Gatekeeper-Position)
- Spezialbeispiel: Der Google-Algorithmus (PageRank) verwendet die *Eigenvector Centrality*

10 Inferenz (ERGMs)

ERGM (Exponential (Family) Random Graph Model)

- ERGMs lassen sich als Regressionsmodelle für Netzwerke vorstellen
- Trivialer Fall: Analyse von Knoten-Merkmalen (*dyadisch unabhängig — Netzwerk hat keinen Effekt*)
- Weitereintwicklung: Analyse von Einfluss der / auf die Netzwerkstruktur
  - Sprich: “Knüpfen Menschen Verbindungen in Abhängigkeit spezifischer Merkmale der Kontakte?”
  - Beispiel: Assortative Mixing / Homophilie / Selectivity & Reinforcement

ERGM-Demo

Analysebeispiel

One can interpret the coefficients of this model in terms of the log-odds of different types of ties: the log-odds of a tie that is completely heterogeneous (the two members differ from each other in race, sex, and grade) is –10.01; the log-odds of a tie that is homogeneous by race only is –8.82 (= –10.01 + 1.20, with rounding error); one that is homogeneous in all three attributes is –4.70 (= 10.01 + 3.23 + 1.20 + 0.88), etc.

Steven M Goodreau, M. S. H. D. R. H. C. T. B. M. M. (2008). A statnet Tutorial. Journal of statistical software, 24(9), 1. NIH Public Access.

11

Resümee & Fragen

Fazit

- Automatisierte Datenerfassung ist einfach und lohnt sich
- Schon einfache Netzwerkanalysen können interessante / unintuitive Ergebnisse bringen
- Methode erfasst bisher “unsichtbare” Aspekte von Online-Kommunikation

Software

- Alle Betriebssysteme: **gephi** (kostenlos)
- Windows: **Pajek** (kostenlos für nichtkommerzielle Nutzung)
- Windows: **UCINET** (Einzellizenz $40-$250)
- ERGMs: **ERGM**-Paket in R

Literatur

- Grundlagenwerk: Stanley Wassermann & Katherine Faust (1994): Social Network Analysis: Methods and Applications
- John Scott (2000): Social Network Analysis: A Handbook
- Programmierung: Toby Segaran: Programming Collective Intelligence
- Matthew A. Russell: Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites

Ausgelassene Themen

- Community-detection: Gruppenbildung nach Nähe/Vernetztheit im Netzwerk
- Mathematische Netzwerktheorie (“Nullmodelle”, zufällige Netzwerke, Modelle für Netzwerkentstehung, Extremfälle)

Vielen Dank für Ihr Interesse & Ihre Ausdauer!

Bildnachweise

- S21: http://www.flickr.com/photos/dielinkebw/5475544113/sizes/l/in/photostream/
- Facebook-Fotovolumen: http://1000memories.com/blog/94-number-of-photos-ever-taken-digital-and-analog-in-shoebox
- Katze: http://www.flickr.com/photos/22963182@N04/3120703441/sizes/z/in/photostream/