Google/Facebook/Amazon: mathematisches Geheimnis von BIG DATA

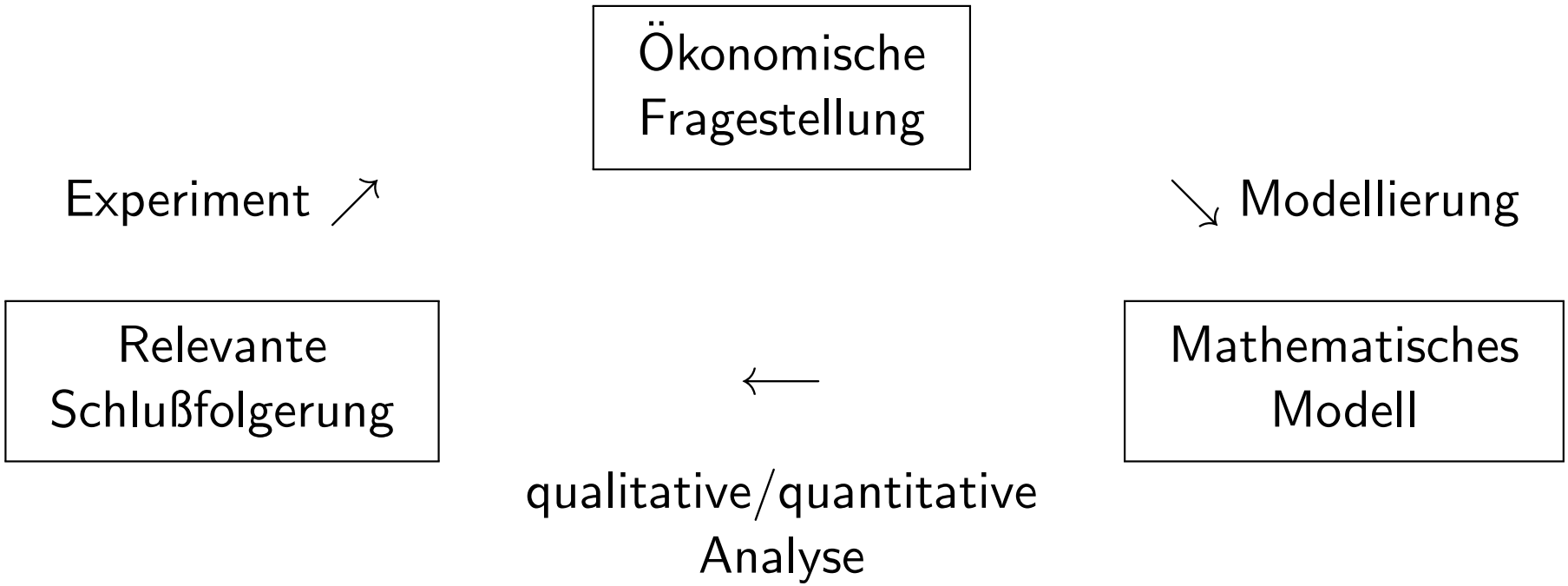
Vladimir Shikhman

Technische Universität Chemnitz

Fakultät für Mathematik

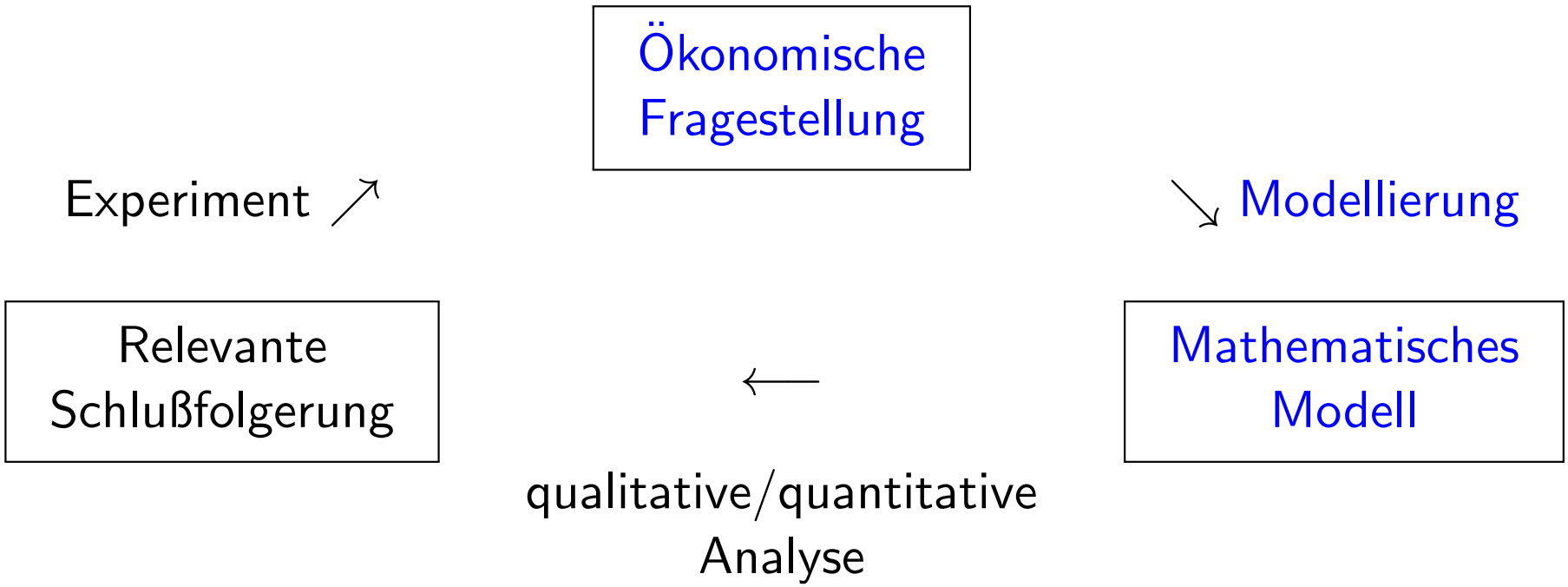
Professur für Wirtschaftsmathematik

WIRTSCHAFTSMATHEMATIK



IMMANUEL KANT, 1796: "Ich behaupte aber, daß in jeder besonderen Naturlehre nur so viel eigentliche Wissenschaft angetroffen werden könne, als darin Mathematik anzutreffen ist."

MATHEMATISCHES MODELL



Ein im Hinblick auf eine bestimmte Fragestellung konstruiertes, vereinfachtes (mathematisches) Abbild eines durch Zusammenhänge zwischen den betrachteten Phänomenen gekennzeichneten Ausschnittes der (ökonomischen) Realität.

BIG DATA in der Ökonomie

Der Sammelbegriff BIG DATA wird für digitale Technologien verwendet, die in technischer Hinsicht für eine neue Ära der Kommunikation und Verarbeitung und in sozialer Hinsicht für einen gesellschaftlichen Umbruch verantwortlich sind. Er steht dabei grundsätzlich fü große digitale Datenmengen, aber auch für deren Analyse, Nutzung, Sammlung, Verwertung und Vermarktung. Das bezeichnende an BIG DATA ist, dass die zu bearbeitenden Datenmengen zu groß, zu komplex, zu schnelllebig oder zu schwach strukturiert sind, um sie mit manuellen und herkömmlichen Methoden der Datenverarbeitung auszuwerten.

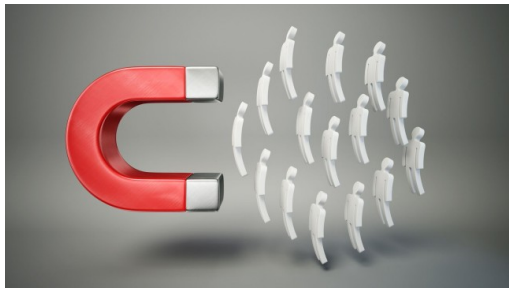
WAS UND WIE LERNEN WIR VON DEN DATEN ?

STANDING ON THE SHOULDERS OF GIANTS

Google

facebook

amazon



RANKING
VON SEITEN

EINFLUSS
VON NUTZERN

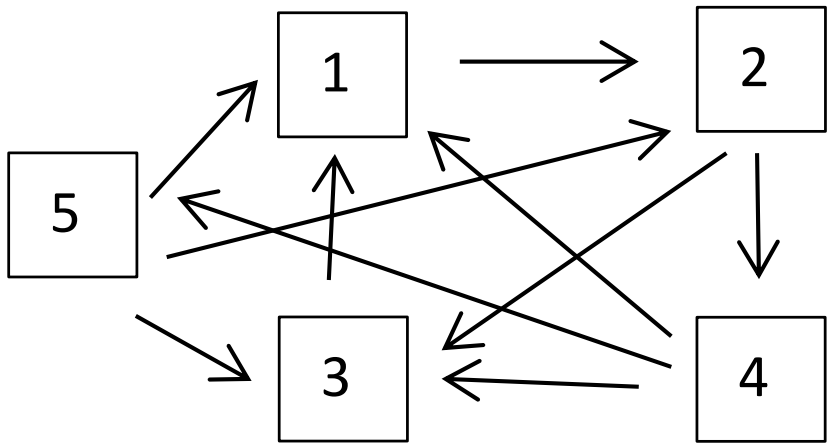
ANTEILE
VON MARKEN

↘ ↓ ↙

MATHEMATISCHES MODELL

GOOGLE NETZWERK

INTERNETSEITEN MIT VERWEISEN / LINKS

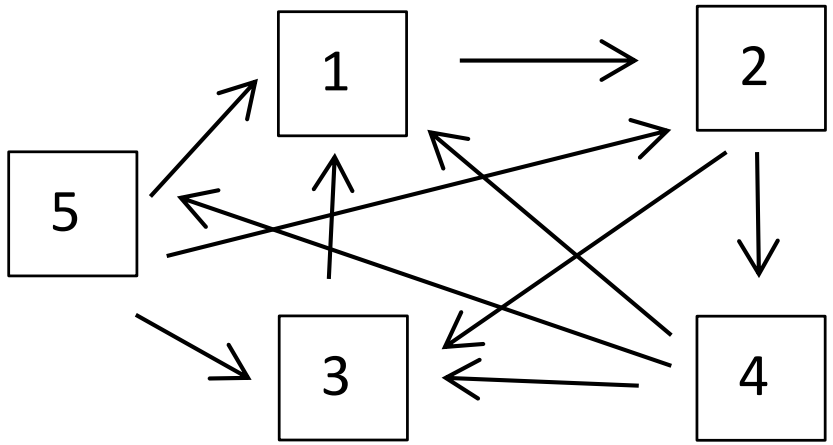


WELCHE SEITEN SIND POPULÄR ?

↓

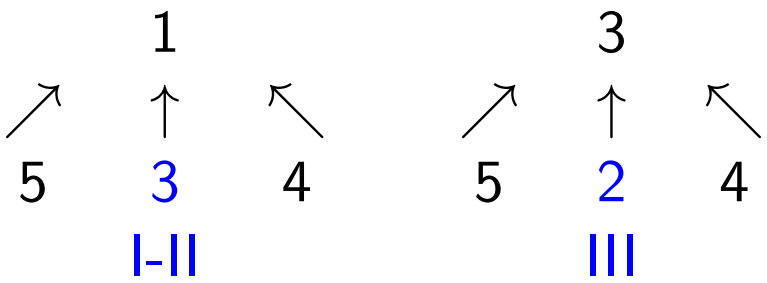
RANKING AUF GOOGLE-ANFRAGE

NAIVER ANSATZ



SEITE	1	2	3	4	5
EINGEHENDE LINKS	3	2	3	1	1
NAIVES RANKING	I-II	III	I-II	IV-V	IV-V

VERGLEICH:

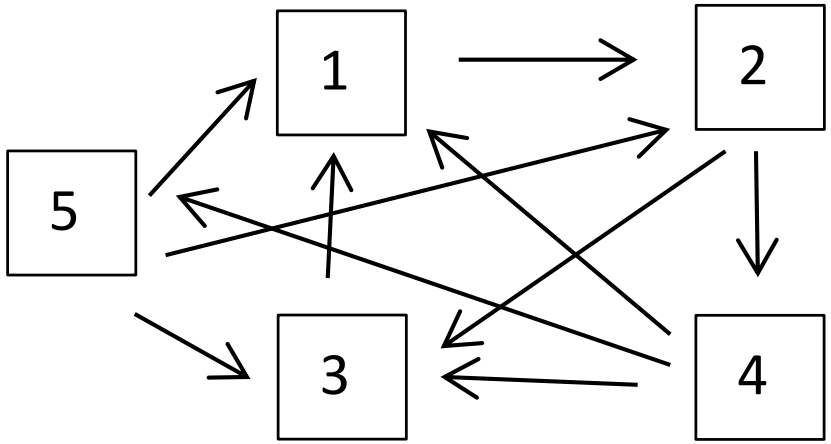


ÜBERGANGSMATRIX

WAHRSCHEINLICHKEIT DES ÜBERGANGES $\boxed{\text{j}} \rightarrow \boxed{\text{i}}$ IST a_{ij}

$i \swarrow j$	1	2	3	4	5
1	0	0	1	1/3	1/3
2	1	0	0	0	1/3
3	0	1/2	0	1/3	1/3
4	0	1/2	0	0	0
5	0	0	0	1/3	0
	1	2	1	3	3

Anzahl ausgehender Links



DATENAUFBEREITUNG : RELATIV VS. ABSOLUT

GOOGLE-PRINZIP

Eine Seite ist populär,
wenn andere populäre Seiten darauf verweisen

$$\underbrace{x_i}_{\substack{\textit{Popularität} \\ \textit{der Seite } i}} = \underbrace{a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n}_{\substack{\textit{Popularität aller Seiten } j, \\ \textit{die mit Wahrscheinichkeit } a_{ij} \textit{ auf } i \textit{ verweisen}}}$$

GLEICHUNGSSYSTEM LÖSEN: $\begin{cases} i = 1, \ldots, n & \text{GLEICHUNGEN} \\ x_1, \ldots, x_n & \text{VARIABLEN} \end{cases}$

GOOGLE-PROBLEM

Übergangsmatrix a_{ij}

$i \swarrow j$	1	2	3	4	5
1	0	0	1	1/3	1/3
2	1	0	0	0	1/3
3	0	1/2	0	1/3	1/3
4	0	1/2	0	0	0
5	0	0	0	1/3	0

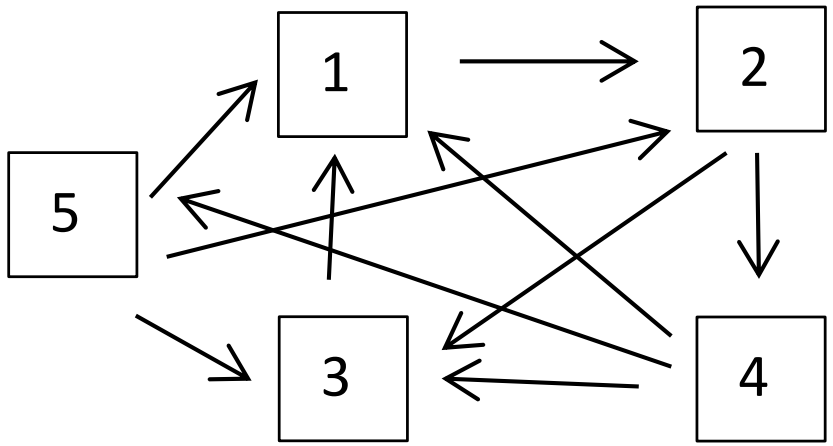
Gleichungssystem

$$x_i = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n$$

$$\begin{aligned} x_1 &= x_3 + \tfrac{1}{3}x_4 + \tfrac{1}{3}x_5 \\ x_2 &= x_1 + \tfrac{1}{3}x_5 \\ x_3 &= \tfrac{1}{2}x_2 + \tfrac{1}{3}x_4 + \tfrac{1}{3}x_5 \\ x_4 &= \tfrac{1}{2}x_2 \\ x_5 &= \tfrac{1}{3}x_4 \end{aligned}$$

SEITE	1	2	3	4	5
GOOGLE POPULARITÄT	17/60	18/60	13/60	9/60	3/60
GOOGLE RANKING	II	I	III	IV	V

VERGLEICH VON RANKINGS



SEITE	1	2	3	4	5
GOOGLE RANKING	II	I	III	IV	V
NAIVES RANKING	I-II	III	I-II	IV-V	IV-V

POPULARITÄT IST AUF SICH SELBST BEZOGEN

FACEBOOK DATEN

NUTZER MIT LIKES

$i \swarrow j$	1	2	3	4	Absoluter Einfluss Anzahl gewonnener Likes
1	0	0	2	1	3
2	5	0	2	1	8
3	0	7	0	1	8
4	4	2	4	0	10

WELCHE NUTZER SIND EINFLUSSREICH ?

$\downarrow$

MANIPULATION FÜR MEINUNGSGESTALTUNG

ZUNEIGUNGSMATRIX

LIKES / ABSOLUT

$i \swarrow j$	1	2	3	4
1	0	0	2	1
2	5	0	2	1
3	0	7	0	1
4	4	2	4	0
	9	9	8	3

Anzahl vergebener Likes

ZUNEIGUNG b_{ij} / RELATIV

$i \swarrow j$	1	2	3	4
1	0	0	2/8	1/3
2	5/9	0	2/8	1/3
3	0	7/9	0	1/3
4	4/9	2/9	4/8	0
	1	1	1	1

Spaltensummen gleich Eins

SOZIALER STATUS

Ein Nutzer ist einflussreich,
wenn andere einflussreiche Nutzer ihm zugeneigt sind

$$\underbrace{y_i}_{\substack{\textit{Einfluss} \\ \textit{des Nutzers } i}} = \underbrace{b_{i1}y_1 + b_{i2}y_2 + \cdots + b_{im}y_m}_{\substack{\textit{Einfluss aller Nutzer } j, \\ \textit{die Zuneigung } b_{ij} \textit{ zu } i \textit{ verspüren}}}$$

GLEICHUNGSSYSTEM LÖSEN: $\begin{cases} i = 1, \ldots, m & \text{GLEICHUNGEN} \\ y_1, \ldots, y_m & \text{VARIABLEN} \end{cases}$

RELATIVER EINFLUSS

Zuneigungsmatrix b_{ij}

$i \swarrow j$	1	2	3	4
1	0	0	2/8	1/3
2	5/9	0	2/8	1/3
3	0	7/9	0	1/3
4	4/9	2/9	4/8	0

Gleichungssystem

$$y_i = b_{i1}y_1 + b_{i2}y_2 + \cdots + b_{im}y_m$$

$$\begin{aligned} y_1 &= \tfrac{2}{8}y_3 + \tfrac{1}{3}y_4 \\ y_2 &= \tfrac{5}{9}y_1 + \tfrac{2}{8}y_3 + \tfrac{1}{3}y_4 \\ y_3 &= \tfrac{7}{9}y_2 + \tfrac{1}{3}y_4 \\ y_4 &= \tfrac{4}{9}y_1 + \tfrac{2}{9}y_2 + \tfrac{4}{8}y_3 \end{aligned}$$

NUTZER	1	2	3	4
RELATIVER EINFLUSS	1	1.55	1.77	1.67
ABSOLUTER EINFLUSS	3	8	8	10

AMAZON DATEN

KAUFVERHALTEN VON KUNDEN

ADIDAS
NIKE
PUMA

Kunde 1	AAAAAAANAPAAA
Kunde 2	PNNNNNNNNA
Kunde 3	NNNNNNNNPAA

WIE ENTWICKELN SICH MARKENANTEILE ?

↓

GRUNDLAGE FÜR MARKTANALYSE

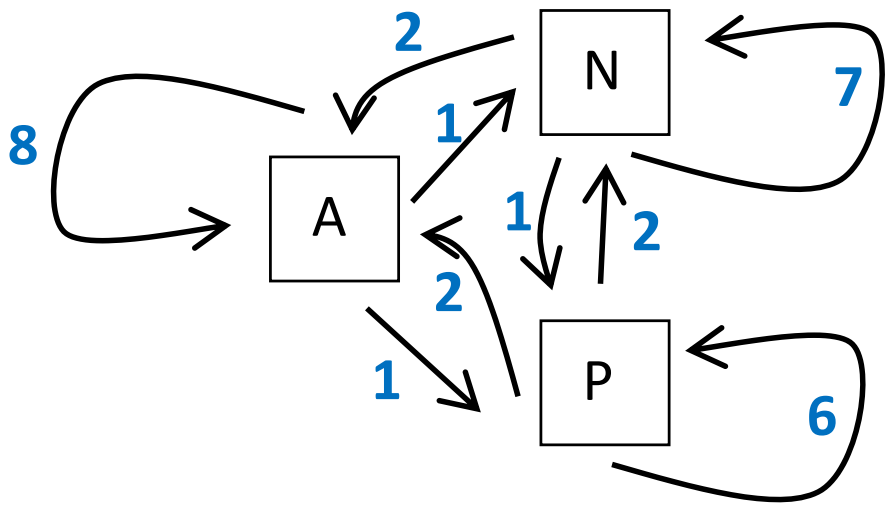
WECHSELGRAPH

KAUFVERHALTEN VON KUNDEN

Kunde 1	AAAAAANAPAAA
Kunde 2	PNNNNNNNNA
Kunde 3	PPPPPPPNPAA

MARKEN	A	N	P
KÄUFE	13	10	10
ANTEIL	40%	30 %	30%

WECHSELWIRKUNGEN MODELLIEREN

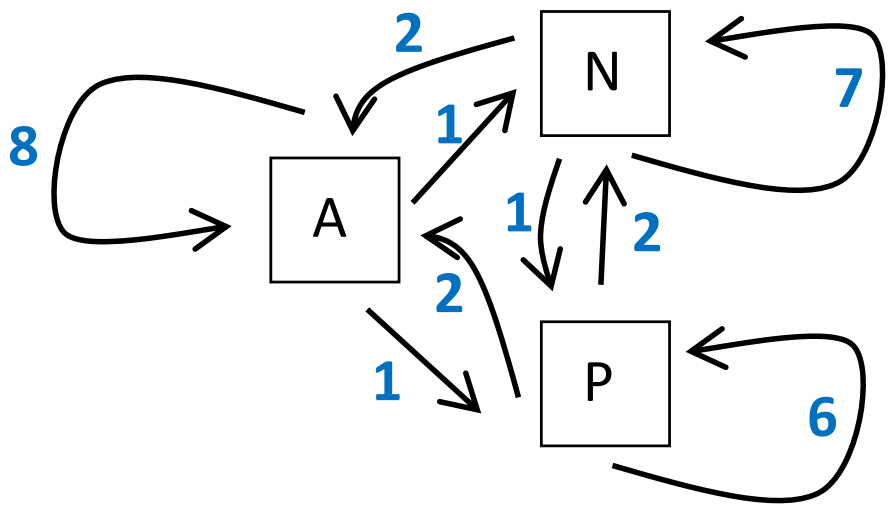


WECHSELMATRIX

WAHRSCHEINLICHKEIT DES WECHSELS j $\rightarrow$ i IST c_{ij}

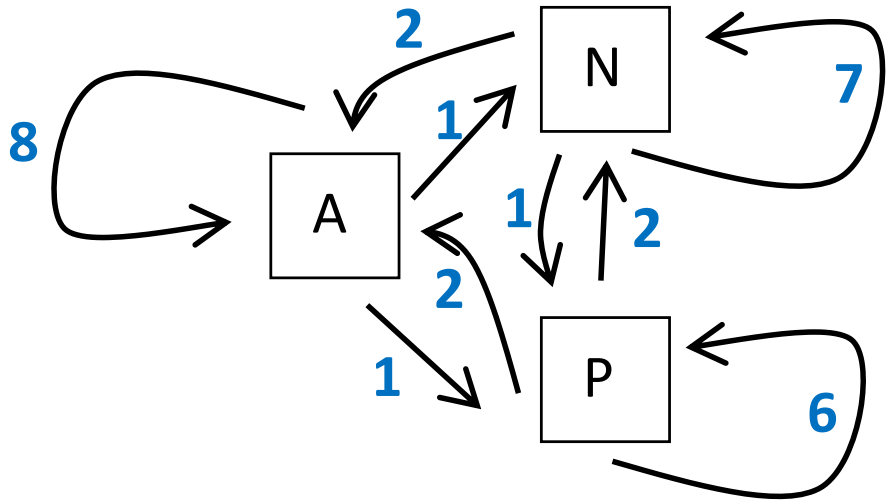
$i \swarrow j$	A	N	P
A	$8/10$	$2/10$	$2/10$
N	$1/10$	$7/10$	$2/10$
P	$1/10$	$1/10$	$6/10$
	10	10	10

Anzahl ausgehender Käufe



WECHSELDYNAMIK

$i \swarrow j$	A	N	P
A	$8/10$	$2/10$	$2/10$
N	$1/10$	$7/10$	$2/10$
P	$1/10$	$1/10$	$6/10$



ANTEIL / ZEIT	t	t+1
A	z_A	$8/10 \cdot z_A + 2/10 \cdot z_N + 2/10 \cdot z_P$
N	z_N	$1/10 \cdot z_A + 7/10 \cdot z_N + 2/10 \cdot z_P$
P	z_P	$1/10 \cdot z_A + 1/10 \cdot z_N + 6/10 \cdot z_P$

STATIONÄRER ANTEIL

UNENDLICHER WECHSEL IN WIEDERHOLUNG

ANTEIL / ZEIT	0	1	2	$\cdots$	∞
A	$1/3$	$8/10 \cdot 1/3 + 2/10 \cdot 1/3 + 2/10 \cdot 1/3 = 12/30$	$132/300$	$\cdots$	50 %
N	$1/3$	$1/10 \cdot 1/3 + 7/10 \cdot 1/3 + 2/10 \cdot 1/3 = 10/30$	$98/300$	$\cdots$	30 %
P	$1/3$	$1/10 \cdot 1/3 + 1/10 \cdot 1/3 + 6/10 \cdot 1/3 = 8/30$	$80/300$	$\cdots$	20 %

STATIONÄRER ANTEIL ÄNDERT SICH NICHT

$$\underbrace{\begin{aligned} z_A &= 8/10 \cdot z_A + 2/10 \cdot z_N + 2/10 \cdot z_P \\ z_N &= 1/10 \cdot z_A + 7/10 \cdot z_N + 2/10 \cdot z_P \\ z_P &= 1/10 \cdot z_A + 1/10 \cdot z_N + 6/10 \cdot z_P \end{aligned}}_{\textit{GOOGLE / FACEBOOK GLEICHUNGSSYSTEM}} \xrightarrow{\text{LÖSEN}} \begin{aligned} z_A^\infty &= 50\% \\ z_N^\infty &= 30\% \\ z_P^\infty &= 20\% \end{aligned}$$

FAZIT

DAS SELBE PRINZIP DER REFLEXIVITÄT

RANKING
VON SEITEN
↑
SURFEN
IM INTERNET

EINFLUSS
VON NUTZERN
↑
AUSTAUSCH
DER ZUNEIGUNG

ANTEILE
VON MARKEN
↑
KÄUFE
VON KUNDEN