Subsampling of Complex Networks

Kai Rathmann

Supervisor: Prof. Dr. Thorsten Strufe

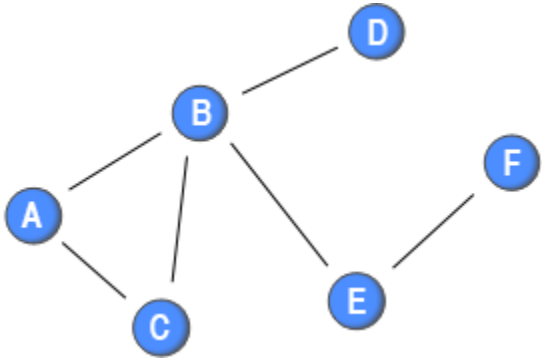
Coordinator: Benjamin Schiller

TECHNISCHE UNIVERSITÄT DARMSTADT

Einleitung

TECHNISCHE UNIVERSITÄT DARMSTADT

- Complex Network G = (V, E)
  - Menge von Knoten V
  - Menge von Kanten E ⊂ (V x V)
    - keine Schleifen
  - zusammenhängend, endlich
- Subsampling
  - Auswahl einer Teilmenge von Knoten V‘ ⊂ V und Kanten E‘ ⊂ E



Motivation

TECHNISCHE UNIVERSITÄT DARMSTADT

- Problemstellung: Großes Netzwerk ist nicht vollständig verwendbar
  - Ressourcenaufwand
  - Zugriffsbeschränkungen
- Gewinn eines repräsentativen Teilnetzwerkes durch Subsampling
- Frage: Welche Eigenschaften des Gesamtnetzwerkes werden durch das Subsampling wie erhalten?

Vorgehen

TECHNISCHE UNIVERSITÄT DARMSTADT

- Gruppierung vieler bekannter Subsampling-Algorithmen
- Versuch einer Klassifikation
- Skizzierung neuer Varianten
- Implementierung in GTNA
- Evaluation

Übersicht

TECHNISCHE UNIVERSITÄT DARMSTADT

- Metriken
- Subsampling-Algorithmen
- Implementierung
- Evaluation
- Zusammenfassung und Fazit

Metriken

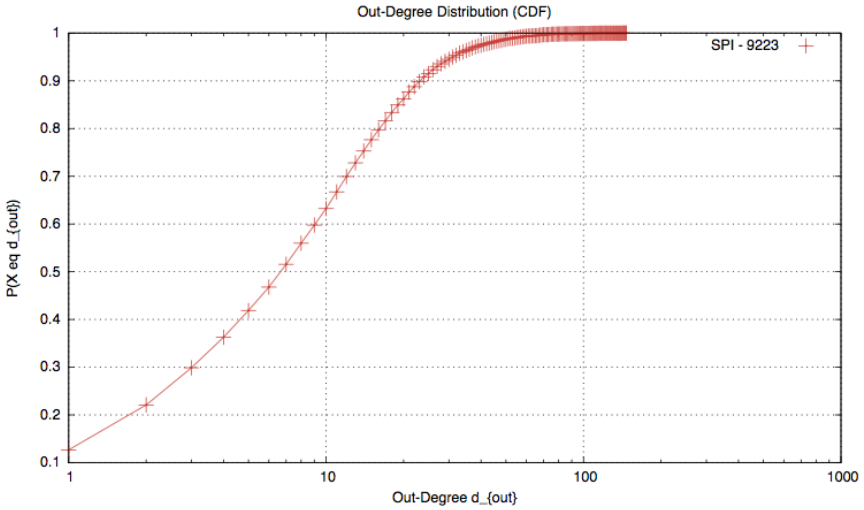
TECHNISCHE UNIVERSITÄT DARMSTADT

- Gradverteilung
- Durchmesser
- Charakteristische Pfadlänge
- Clusterkoeffizient
- Netzwerkgröße

Metriken: Gradverteilung

TECHNISCHE UNIVERSITÄT DARMSTADT

- Grad k_v eines Knoten v: Anzahl Kanten, die inzident zu diesem Knoten sind
  - Ausgangsgrad k^+_v: ausgehende Kanten (v, w)
  - Eingangsgrad k^-_v: eingehende Kanten (w, v)
- Gradverteilung: Verteilung von Knotengraden
- Skalenfreie Netzwerke
  - $P(k) \sim k^{-\gamma}$



Metriken: Durchmesser, Char. Pfadlänge

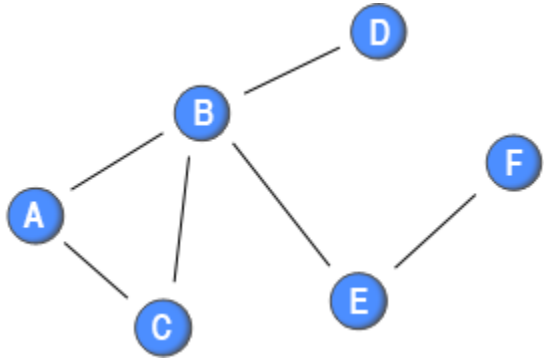
TECHNISCHE UNIVERSITÄT DARMSTADT

- Pfad (v_1, ..., v_n): Knoten v_1, ... v_n, so dass Kanten (v_1, v_2), ..., (v_{n-1}, v_n) existieren
- Länge eines Pfades (v_1, ..., v_n): n – 1
- Abstand zweier Knoten: Länge des kürzesten Pfades
- Durchmesser: Maximaler Abstand zweier Knoten im Netzwerk
- Charakteristische Pfadlänge: Durchschnittlicher Abstand zweier Knoten im Netzwerk
- Milgram‘s Experiment
- Six Degrees of Separation

Metriken: Clusterkoeffizient

TECHNISCHE UNIVERSITÄT DARMSTADT

- Triplet:
  - drei Knoten, die durch zwei (offenes Triplet) oder drei (geschlossenes Triplet) Kanten verbunden sind
- Clusterkoeffizient:
  - Anzahl geschlossener Triplets / Anzahl aller Triplets



Metriken: Netzwerkgröße

TECHNISCHE UNIVERSITÄT DARMSTADT

- Anzahl Knoten im Netzwerk
- relevant nur für *wiederkehrende* Subsampling-Algorithmen

Subsampling-Algorithmen

TECHNISCHE UNIVERSITÄT DARMSTADT

- Uniform Node Sampling
- Classic Random Walk
- Metropolis-Hastings Random Walk
- Frontier Sampling
- Breadth-First Search
- Depth-First Search
- Forest Fire
- Snowball Sampling
- Respondent-Driven Sampling

Algorithmen: Uniform Node Sampling

TECHNISCHE UNIVERSITÄT DARMSTADT

- Uniforme Auswahl von Knoten

Algorithmen: Classic Random Walk

TECHNISCHE UNIVERSITÄT DARMSTADT

- Uniforme Wahl des Startknotens
- In jedem Schritt
  - Uniforme Wahl eines Nachbarknotens
    - $p(v, w) = 1/k_v$
- Wiederholte Wahl von Knoten möglich

Algorithmen: Metropolis-Hastings R.W.

TECHNISCHE UNIVERSITÄT DARMSTADT

- Uniforme Wahl des Startknotens
- In jedem Schritt
  - Gewichtete Wahl eines Nachbarknotens
    - $p(v, w) = 1/k_v * \min(1, k_v/k_w)$ falls w Nachbar von v
    - $p(v, v) = 1 - \Sigma_{w \neq v} p(v, w)$
- Wiederholte Wahl von Knoten möglich

Algorithmen: Frontier Sampling

TECHNISCHE UNIVERSITÄT DARMSTADT

- m Instanzen eines klassischen Random Walks
- In jedem Schritt wähle einen Random Walk aus, um einen Schritt auszuführen
  - $p(i) = k(v_i) / \Sigma_i k(v_i)$

Algorithmen: Breadth-First Search

TECHNISCHE UNIVERSITÄT DARMSTADT

- Markiere den Startknoten
- Speichere den Startknoten in der Schlange
- Solange die Schlange nicht leer ist
  - Entnimm den ersten Knoten v aus der Schlange
  - Für alle (ausgehenden) Kanten (v, w)
    - Falls w nicht markiert ist
      - Markiere w
      - Speichere w in der Schlange

Algorithmen: Depth-First Search

TECHNISCHE UNIVERSITÄT DARMSTADT

- Markiere v
- Für alle (ausgehenden) Kanten (v, w)
  - Falls w unmarkiert ist
    - Depth-First Search(w)

Algorithmen: Forest Fire

TECHNISCHE UNIVERSITÄT DARMSTADT

- Ähnlich Breadth-First Search
- „Markiere w“ für alle unmarkierten Nachbarknoten nur mit Wahrscheinlichkeit p
  - p=1: BFS

Algorithmen: Snowball Sampling

TECHNISCHE UNIVERSITÄT DARMSTADT

- Markiere den Startknoten
- Speichere den Startknoten in der Schlange
- Solange die Schlange nicht leer ist
  - Entnimm den ersten Knoten v aus der Schlange
    - Für max. n (ausgehende) Kanten (v, w) mit unmarkiertem w
      - Markiere w
      - Speichere w in der Schlange

Algorithmen: Respondent-Driven Sampling

TECHNISCHE UNIVERSITÄT DARMSTADT

- Ähnlich Snowball Sampling
- Wiederholte Wahl von Knoten möglich

Algorithmen: Klassifikation

TECHNISCHE UNIVERSITÄT DARMSTADT

Algorithmus	Nachbarn
UNS	keine
RW	1
MHRW	1
FS	1
BFS	alle
DFS	1
FF	alle (mit p)
SS	mehrere
RDS	mehrere

Algorithmen: Klassifikation

TECHNISCHE UNIVERSITÄT DARMSTADT

Nachbarn	Algorithmus
n = 0	UNS
n = 1	RW, MHRW, FS, DFS
1 < n < ∞	SS, RDS
n = ∞	BFS, FF

Algorithmen: Klassifikation

TECHNISCHE UNIVERSITÄT DARMSTADT

Nachbarn	nicht wiederkehrend	wiederkehrend
n = 0	UNS	
n = 1	DFS	RW, MHRW, FS
1 < n < ∞	SS	RDS
n = ∞	BFS, FF	

Algorithmen: Klassifikation

TECHNISCHE UNIVERSITÄT DARMSTADT

Nachbarn		nicht wiederkehrend	wiederkehrend
n = 0		UNS	
n = 1	geordnet	DFS	
	uniform		RW, FS
	gewichtet		MHRW
1 < n < ∞	uniform	SS	RDS
n = ∞	alle	BFS	
	p fest	FF	

Algorithmen: Klassifikation

TECHNISCHE UNIVERSITÄT DARMSTADT

Nachbarn		nicht wieder-kehrend	wiederkehrend	mehr-dimensional
n = 0		UNS		
n = 1	geordnet	DFS		
	uniform		RW	FS
	gewichtet		MHRW	
1 < n < ∞	uniform	SS	RDS	
n = ∞	alle	BFS		
	p fest	FF		

Algorithmen: Klassifikation

TECHNISCHE UNIVERSITÄT DARMSTADT

Nachbarn		nicht wieder-kehrend	wieder-kehrend	unabhängig mehr-dimensional	abhängig mehr-dimensional
n = 0		UNS			
n = 1	geordnet	DFS			
	uniform		RW		FS
	gewichtet		MHRW		
1 < n < ∞	geordnet				
	uniform	SS	RDS		
	gewichtet				
n = ∞	alle	BFS			
	p fest	FF			
	p gewichtet				

Algorithmen: Klassifikation

TECHNISCHE UNIVERSITÄT DARMSTADT

Nachbarn		nicht wieder-kehrend	wieder-kehrend	unabhängig mehr-dimensional	abhängig mehr-dimensional
n = 0		UNS			
n = 1	geordnet	DFS			
	uniform	*	RW	*	FS
	gewichtet	*	MHRW	*	*
1 < n < ∞	geordnet	*			
	uniform	SS	RDS	*	*
	gewichtet	*	*	*	*
n = ∞	alle	BFS			
	p fest	FF			
	p gewichtet	*			

Algorithmen: Klassifikation

TECHNISCHE UNIVERSITÄT DARMSTADT

Nachbarn		nicht wieder-kehrend	wieder-kehrend	unabhängig mehr-dimensional	abhängig mehr-dimensional
n = 0		UNS			
n = 1	geordnet	$DFS_{geordnet}$			
	uniform	$DFS_{uniform}$	$RW_{uniform}$	$mRW_{uniform}$	$FS_{uniform}$
	gewichtet	$DFS_{gewichtet}$	$RW_{gewichtet}$	$mRW_{gewichtet}$	$FS_{gewichtet}$
1 < n < ∞	geordnet	$SS_{geordnet}$			
	uniform	$SS_{uniform}$	$RDS_{uniform}$	$mRDS_{uniform}$	$FS+_{uniform}$
	gewichtet	$SS_{gewichtet}$	$RDS_{gewichtet}$	$mRDS_{gewicht}$	$FS+_{gewichtet}$
n = ∞	alle	BFS			
	p fest	FF_{fest}			
	p gewichtet	$FF_{gewichtet}$			

Implementierung

TECHNISCHE UNIVERSITÄT DARMSTADT

- `Transformation`
  - `isApplicable(Graph): boolean`
  - `transform(Graph): Graph`
- `Subsampling`
  - `subsample(Graph): Collection<Node>`
  - Normalisierung
- Subklassen für alle Algorithmen

GTNA

Graph-Theoretic Network Analyzer

Evaluation

TECHNISCHE UNIVERSITÄT DARMSTADT

- Setup
- Netzwerke
- Ergebnisse

Evaluation: Setup

TECHNISCHE UNIVERSITÄT DARMSTADT

- 100 Durchläufe für jede Kombination aus Netzwerktyp und Subsampling-Algorithmus
  - Netzwerkmodelle: 1 Durchlauf pro generiertem Netzwerk
  - Reale Netzwerke: 100 Durchläufe
- 10% Budget
- n = 3
- p = 0,7
- m = 10

Evaluation: Netzwerke

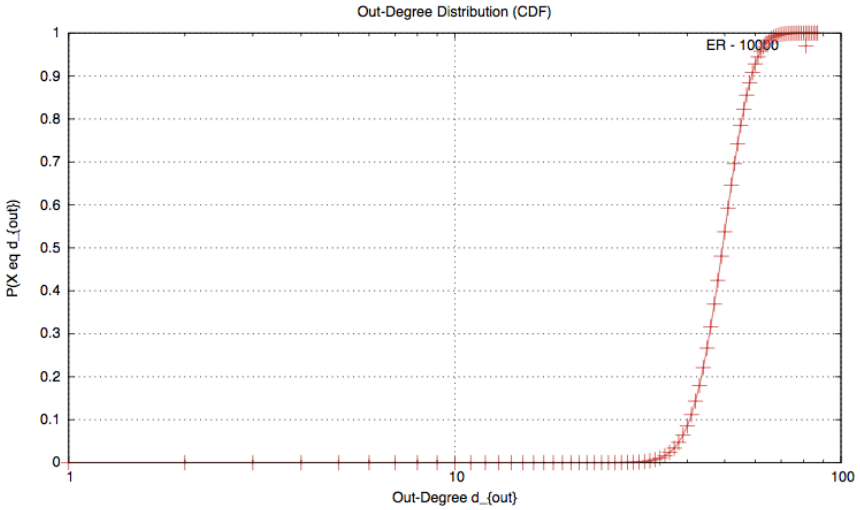
TECHNISCHE UNIVERSITÄT DARMSTADT

- Erdős–Rényi
- Barabási–Albert
- Studentenportal Ilmenau
- Web of Trust (ungerichtet)
- Web of Trust (gerichtet)

Eval. – Netzwerke: Erdős–Rényi

TECHNISCHE UNIVERSITÄT DARMSTADT

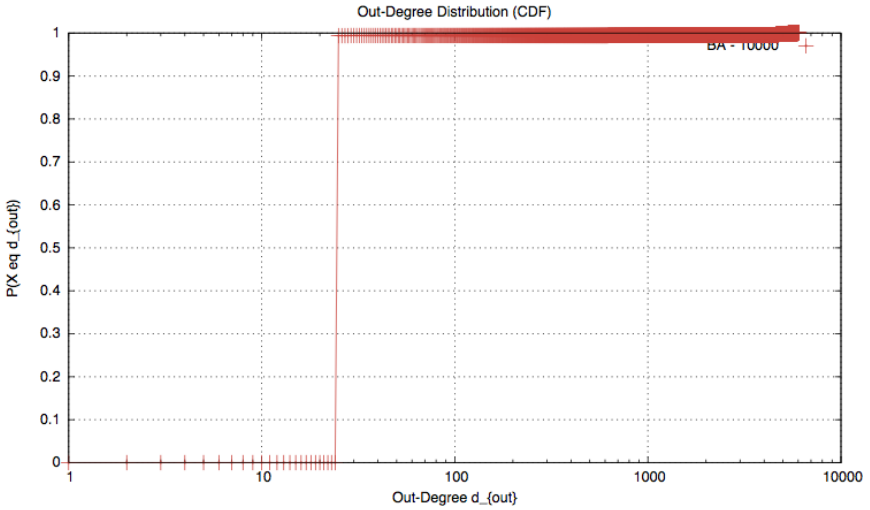
- Setzt Kanten zwischen Knotenpaaren mit gleicher Wahrscheinlichkeit, unabhängig von anderen Kanten
- 10.000 Knoten
- Durchschnittsgrad 50
- Durchschnittswerte:
  - Durchmesser 4,0
  - Ch. Pfadlänge 2,77
  - Clusterkoeffizient 0.0050



Eval. – Netzwerke: Barabási–Albert

TECHNISCHE UNIVERSITÄT DARMSTADT

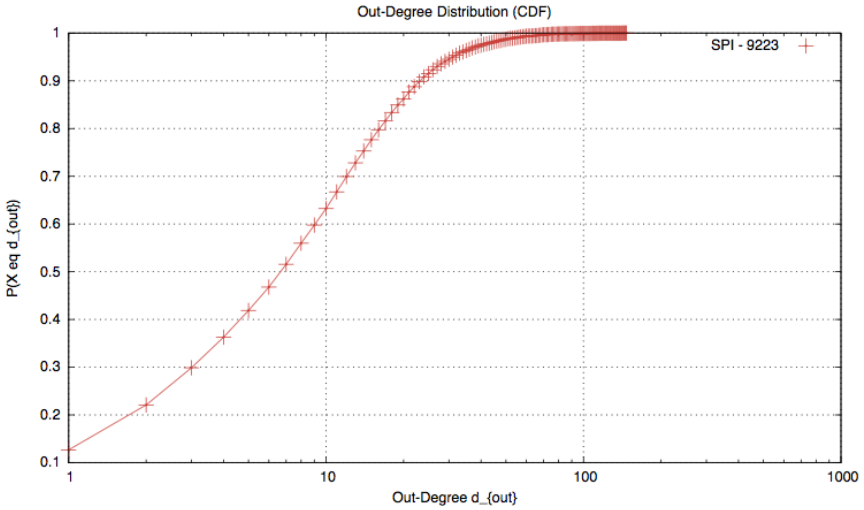
- Preferential attachment
- Skalenfreies Netzwerk
- 10.000 Knoten
- Kantenzahl pro Knoten 25
- Durchschnittswerte:
  - Durchmesser 2,3
  - Ch. Pfadlänge 2,00
  - Clusterkoeffizient 0,4592



Eval. – Netzwerke: Studentenportal Ilmenau

TECHNISCHE UNIVERSITÄT DARMSTADT

- Soziales Netzwerk
- 9.223 Knoten
- Durchmesser 12
- Ch. Pfadlänge 4,66
- Clusterkoeffizient 0,2964



Eval. – Netzwerke: Web of Trust

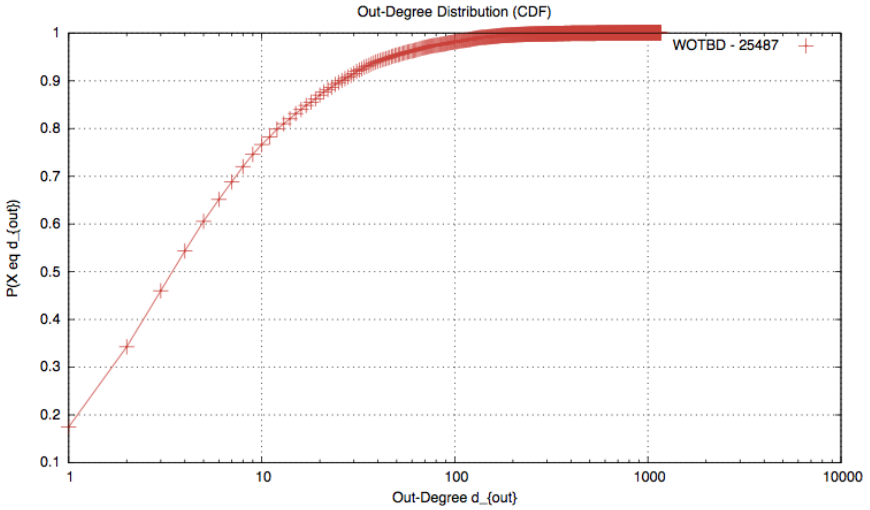
TECHNISCHE UNIVERSITÄT DARMSTADT

- Zertifizierungsgraph von PGP
- Evaluation sowohl der ungerichteten als auch der gerichteten Variante
- 25.487 Knoten

Eval. – Netzwerke: Web of Trust (ungerichtet)

TECHNISCHE UNIVERSITÄT DARMSTADT

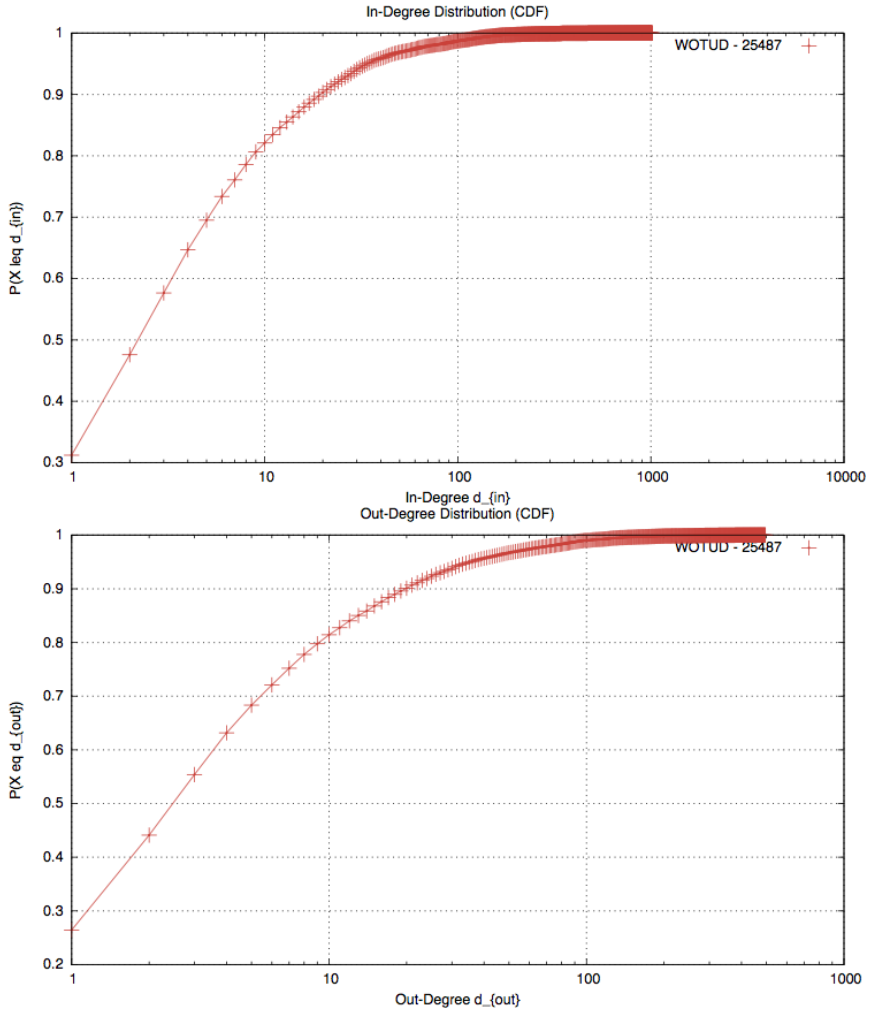
- Durchmesser 15
- Ch. Pfadlänge 5,07
- Clusterkoeffizient 0,4665



Eval. – Netzwerke: Web of Trust (gerichtet)

TECHNISCHE UNIVERSITÄT DARMSTADT

- Durchmesser 25
- Ch. Pfadlänge 5,99
- Clusterkoeffizient 0,3707



Evaluation: Ergebnisse

TECHNISCHE UNIVERSITÄT DARMSTADT

- Netzwerkgröße
- Gradverteilung
- Durchmesser
- Charakteristische Pfadlänge
- Clusterkoeffizient

Eval. – Ergebnisse: Netzwerkgröße

TECHNISCHE UNIVERSITÄT DARMSTADT

Nachbarn		nicht wieder-kehrend	wieder-kehrend	unabhängig mehr-dimensional	abhängig mehr-dimensional
n = 0		UNS			
n = 1	geordnet	$DFS_{geordnet}$			
	uniform	$DFS_{uniform}$	$RW_{uniform}$	$mRW_{uniform}$	$FS_{uniform}$
	gewichtet	$DFS_{gewichtet}$	$RW_{gewichtet}$	$mRW_{gewichtet}$	$FS_{gewichtet}$
1 < n < ∞	geordnet	$SS_{geordnet}$			
	uniform	$SS_{uniform}$	$RDS_{uniform}$	$mRDS_{uniform}$	$FS+_{uniform}$
	gewichtet	$SS_{gewichtet}$	$RDS_{gewichtet}$	$mRDS_{gewicht}$	$FS+_{gewichtet}$
n = ∞	alle	BFS			
	p fest	FF_{fest}			
	p gewichtet	$FF_{gewichtet}$			

Eval. – Ergebnisse: Netzwerkgröße

TECHNISCHE UNIVERSITÄT DARMSTADT

Nachbarn		nicht wieder-kehrend	wieder-kehrend	unabhängig mehr-dimensional	abhängig mehr-dimensional
n = 0					
n = 1	geordnet				
	uniform		$RW_{uniform}$	$mRW_{uniform}$	$FS_{uniform}$
	gewichtet		$RW_{gewichtet}$	$mRW_{gewichtet}$	$FS_{gewichtet}$
1 < n < ∞	geordnet				
	uniform		$RDS_{uniform}$	$mRDS_{uniform}$	$FS+_{uniform}$
	gewichtet		$RDS_{gewichtet}$	$mRDS_{gewicht}$	$FS+_{gewichtet}$
n = ∞	alle				
	p fest				
	p gewichtet				

Eval. – Ergebnisse: Netzwerkgröße

TECHNISCHE UNIVERSITÄT DARMSTADT

- Reale Netzwerke
  - Primärer Einfluss: Nachbarstrategie
    1. n = 1 uniform
    2. 1 < n < ∞ gewichtet
    3. 1 < n < ∞ uniform
    4. n = 1 gewichtet
  - Sekundärer Einfluss: Dimensionale Strategie
    1. abhängig mehrdimensional
    2. unabhängig mehrdimensional
    3. eindimensional

Eval. – Ergebnisse: Netzwerkgröße

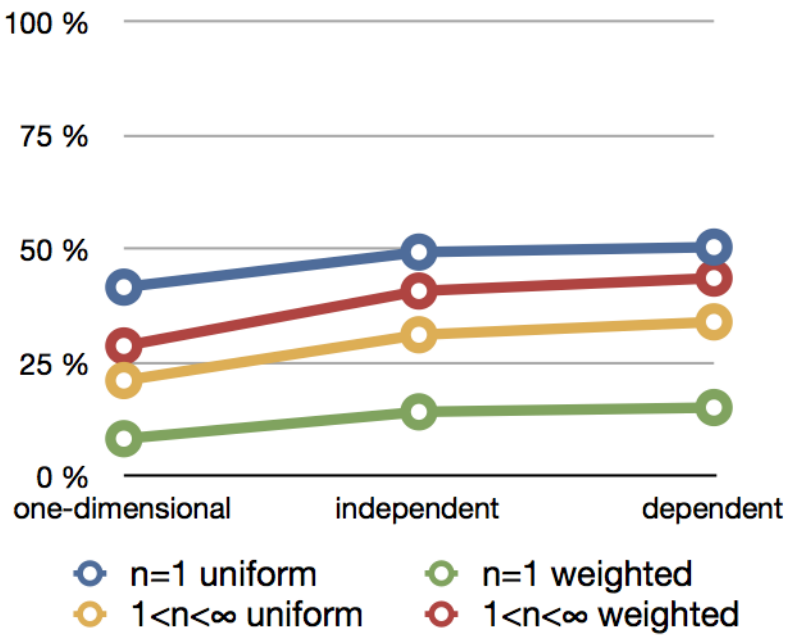
TECHNISCHE UNIVERSITÄT DARMSTADT

Nachbarn		nicht wieder-kehrend	wieder-kehrend	unabhängig mehr-dimensional	abhängig mehr-dimensional
n = 0					
n = 1	geordnet				
	uniform		$RW_{uniform}$	$mRW_{uniform}$	$FS_{uniform}$
	gewichtet		$RW_{gewichtet}$	$mRW_{gewichtet}$	$FS_{gewichtet}$
1 < n < ∞	geordnet				
	uniform		$RDS_{uniform}$	$mRDS_{uniform}$	$FS+_{uniform}$
	gewichtet		$RDS_{gewichtet}$	$mRDS_{gewicht}$	$FS+_{gewichtet}$
n = ∞	alle				
	p fest				
	p gewichtet				

Eval. – Ergebnisse: Netzwerkgröße

TECHNISCHE UNIVERSITÄT DARMSTADT

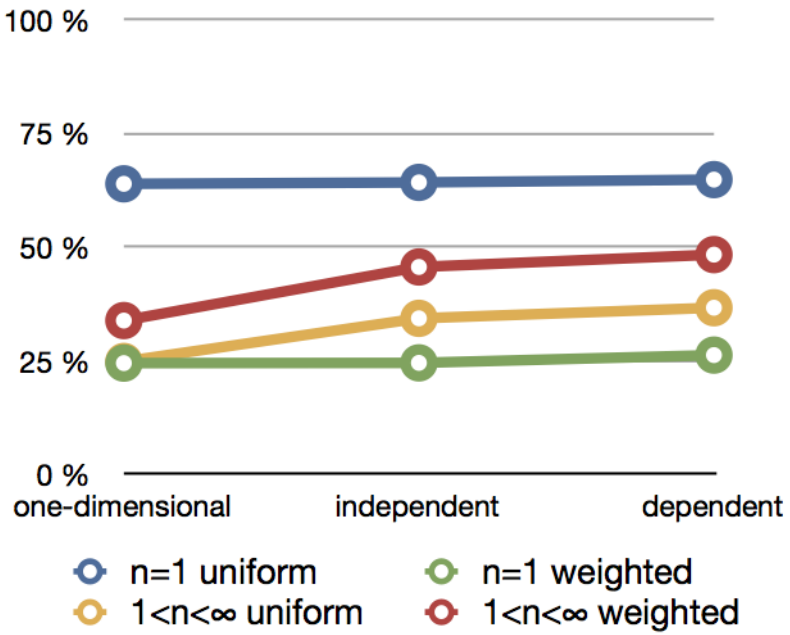
Web of Trust (gerichtet)



Eval. – Ergebnisse: Netzwerkgröße

TECHNISCHE UNIVERSITÄT DARMSTADT

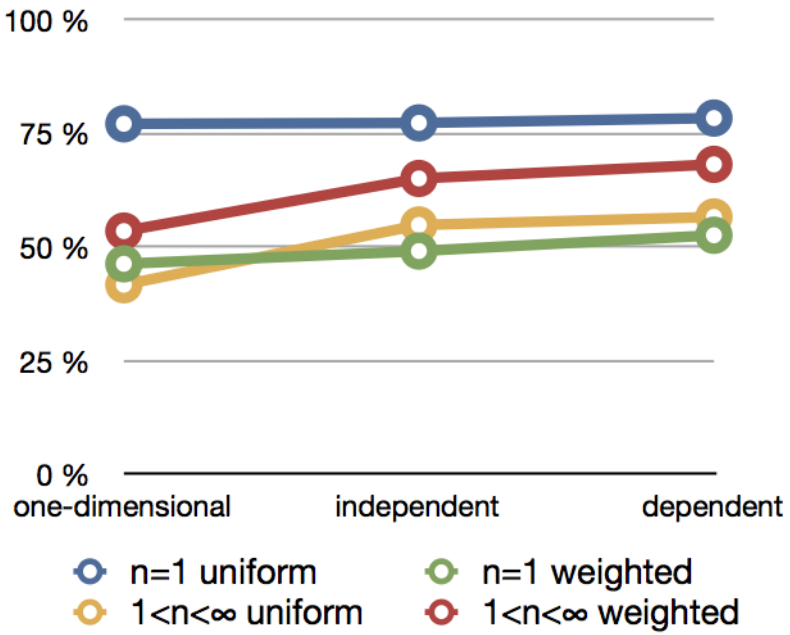
Web of Trust (ungerichtet)



Eval. – Ergebnisse: Netzwerkgröße

TECHNISCHE UNIVERSITÄT DARMSTADT

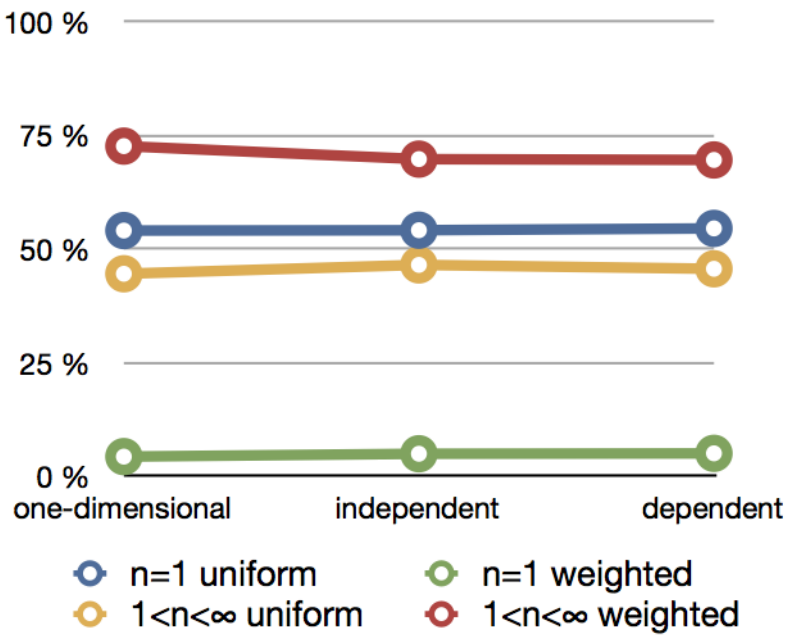
Studentenportal Ilmenau



Eval. – Ergebnisse: Netzwerkgröße

TECHNISCHE UNIVERSITÄT DARMSTADT

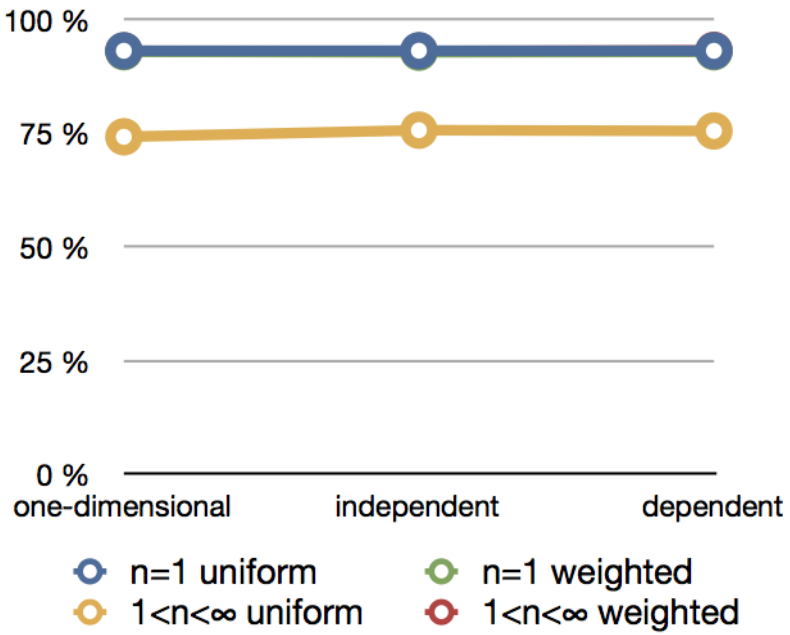
Barabási–Albert



Eval. – Ergebnisse: Netzwerkgröße

TECHNISCHE UNIVERSITÄT DARMSTADT

Erdős–Rényi



Eval. – Ergebnisse: Gradverteilung

TECHNISCHE UNIVERSITÄT DARMSTADT

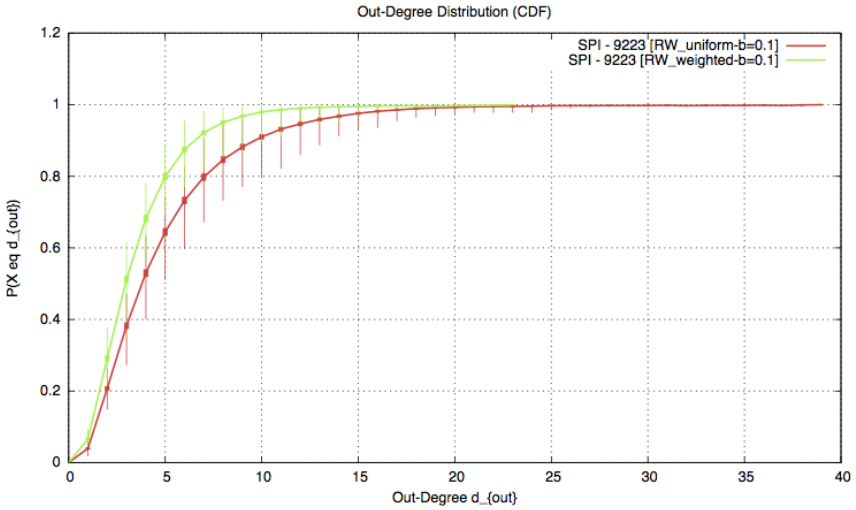
- Ergebnisse zu $RW_{gewichtet}$ und $FS_{uniform}$ reproduziert
- Effekt von Gewichtung gezeigt für SS, mRW, FS, RDS, mRDS, FS+, FF
- $DFS_{uniform} \approx DFS_{gewichtet}$
- Konzeptkombination $FS_{gewichtet}$ verstärkt Effekt

Eval. – Ergebnisse: Gradverteilung

TECHNISCHE UNIVERSITÄT DARMSTADT

$RW_{uniform}$ vs. $RW_{gewichtet}$

am Beispiel Studentenportal Ilmenau

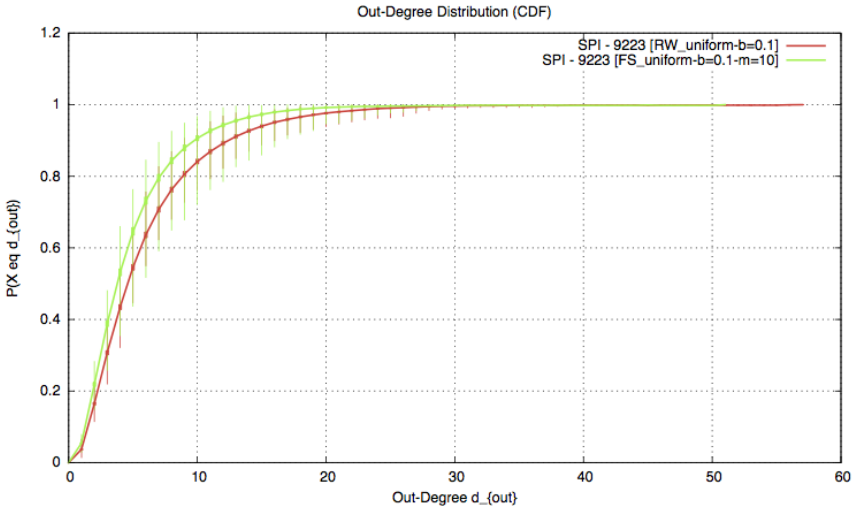


Eval. – Ergebnisse: Gradverteilung

TECHNISCHE UNIVERSITÄT DARMSTADT

$RW_{uniform}$ vs. $FS_{uniform}$

am Beispiel Studentenportal Ilmenau

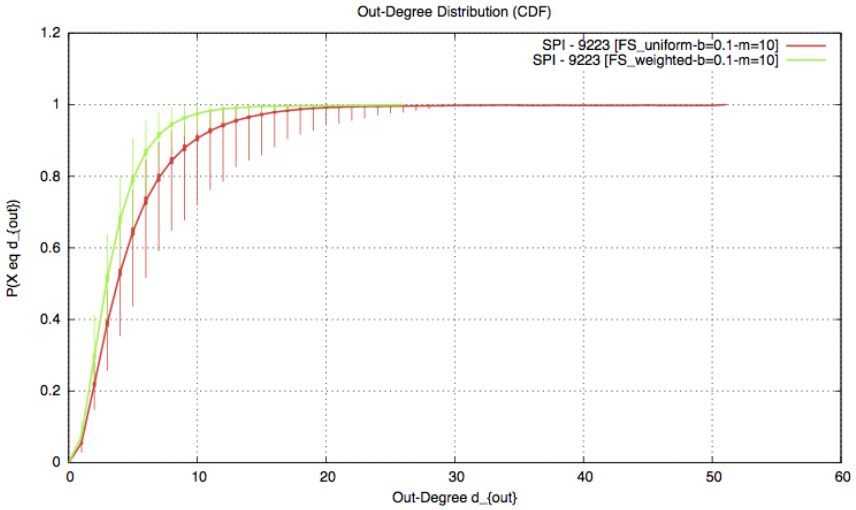


Eval. – Ergebnisse: Gradverteilung

TECHNISCHE UNIVERSITÄT DARMSTADT

$FS_{uniform}$ vs. $FS_{gewichtet}$

am Beispiel Studentenportal Ilmenau

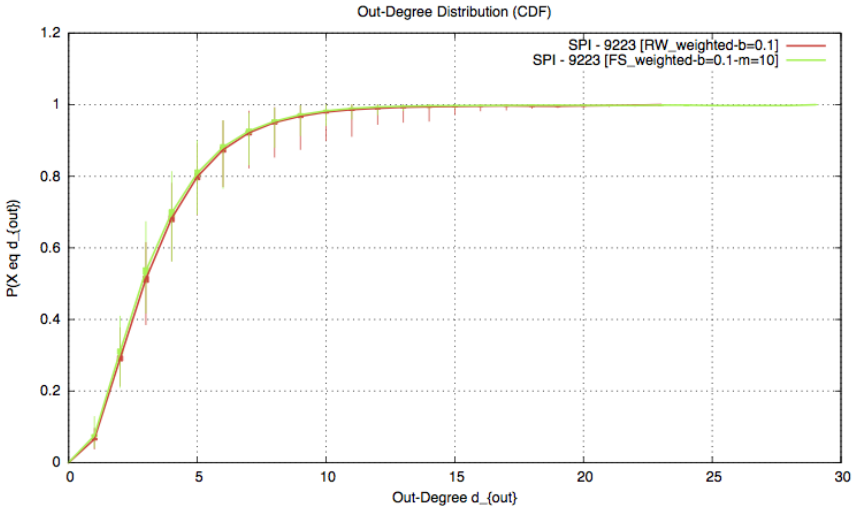


Eval. – Ergebnisse: Gradverteilung

TECHNISCHE UNIVERSITÄT DARMSTADT

$RW_{gewichtet}$ vs. $FS_{gewichtet}$

am Beispiel Studentenportal Ilmenau



Eval. – Ergebnisse: Durchmesser und Ch. Pf.

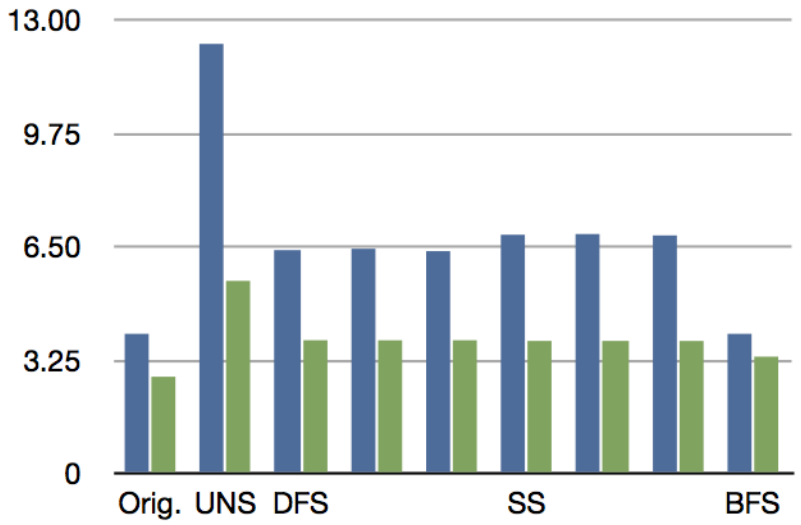
TECHNISCHE UNIVERSITÄT DARMSTADT

- DFS-Varianten insgesamt am originaltreuesten
- SS-Varianten produzieren niedrigere Werte als DFS-Varianten in allen evaluierten Netzwerken außer Erdős–Rényi
- UNS produziert höchste Werte in den ungerichteten Netzwerken, sehr niedrige in dem gerichteten Netzwerk
- BFS produziert niedrigste Werte in allen evaluierten Netzwerken außer Barabási–Albert

Eval. – Ergebnisse: Durchmesser und Ch. Pf.

TECHNISCHE UNIVERSITÄT DARMSTADT

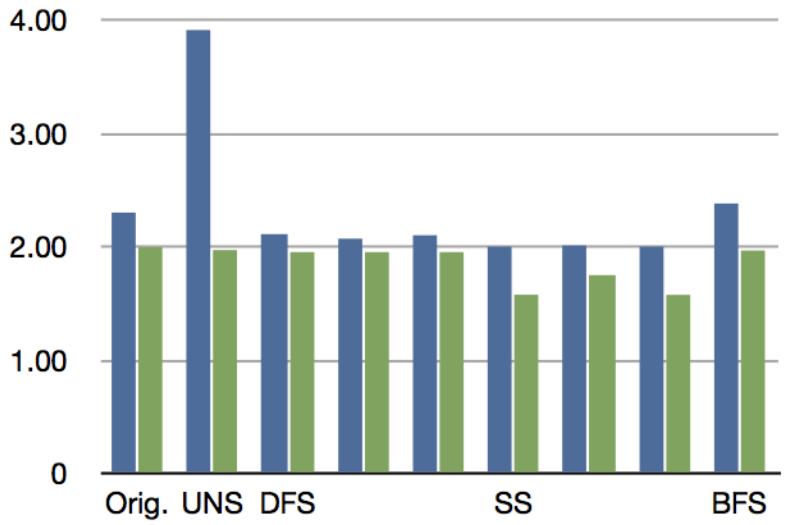
Erdős–Rényi



Eval. – Ergebnisse: Durchmesser und Ch. Pf.

TECHNISCHE UNIVERSITÄT DARMSTADT

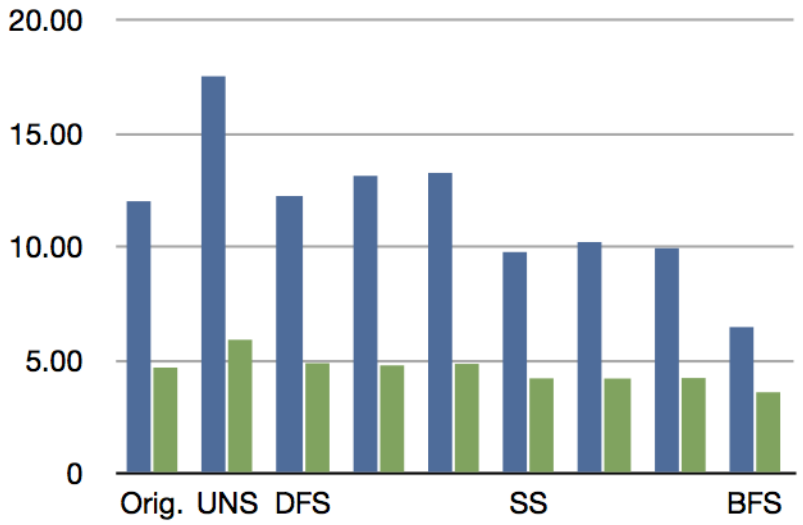
Barabási–Albert



Eval. – Ergebnisse: Durchmesser und Ch. Pf.

TECHNISCHE UNIVERSITÄT DARMSTADT

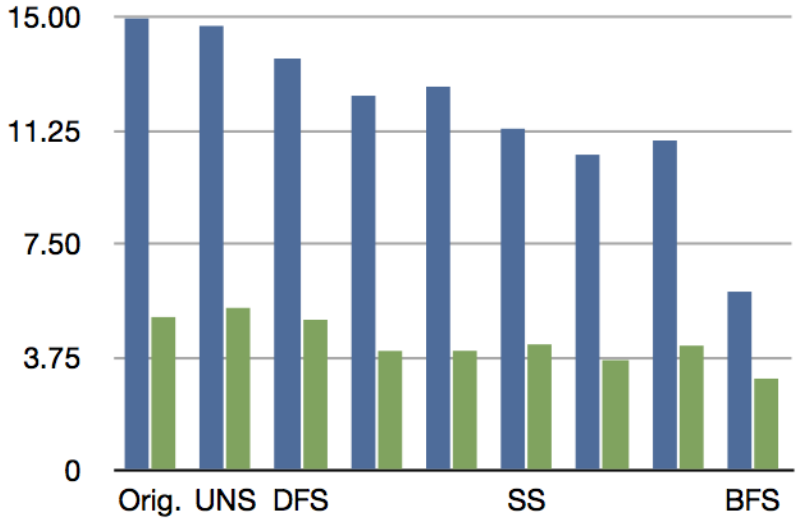
Studentenportal Ilmenau



Eval. – Ergebnisse: Durchmesser und Ch. Pf.

TECHNISCHE UNIVERSITÄT DARMSTADT

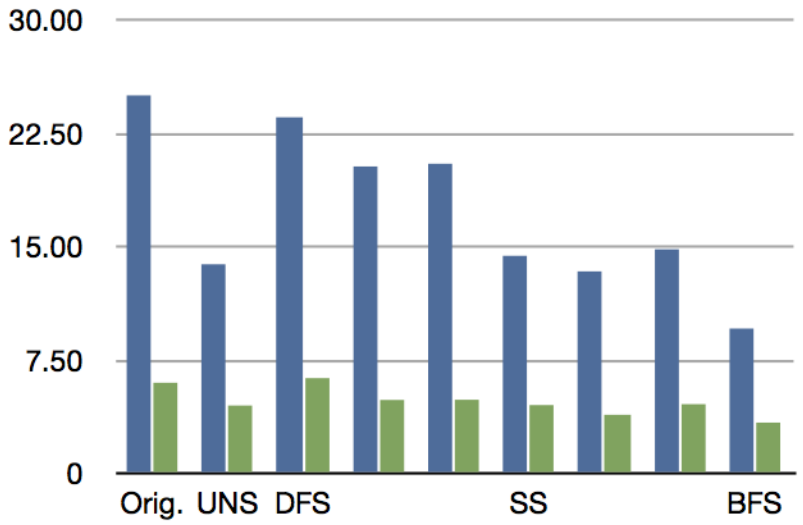
Web of Trust (ungerichtet)



Eval. – Ergebnisse: Durchmesser und Ch. Pf.

TECHNISCHE UNIVERSITÄT DARMSTADT

Web of Trust (gerichtet)



Eval. – Ergebnisse: Clusterkoeffizient

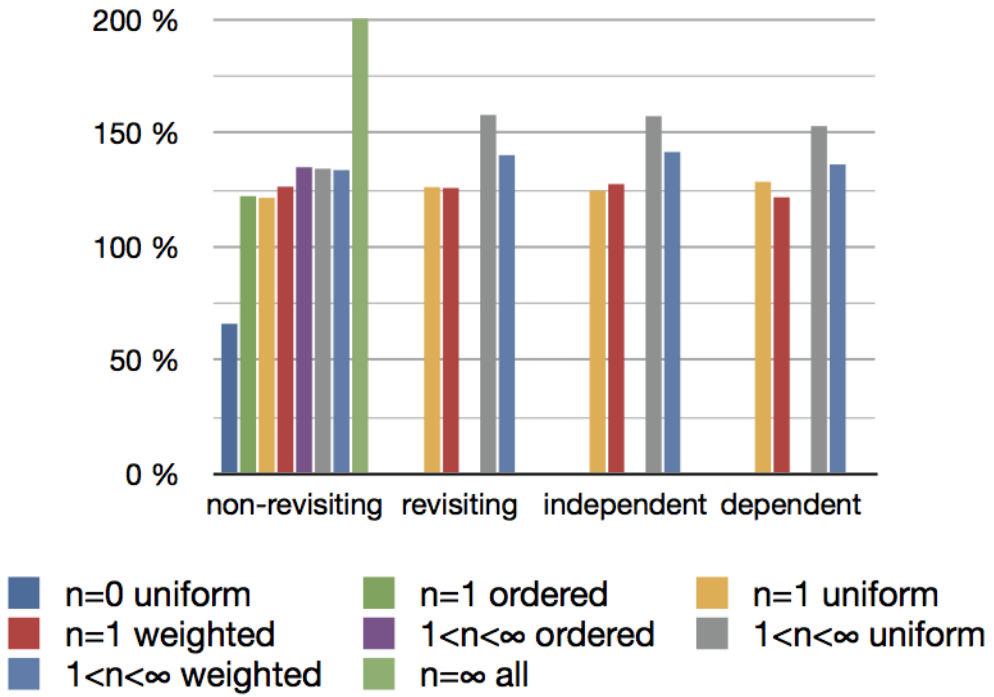
TECHNISCHE UNIVERSITÄT DARMSTADT

- Algorithmen mit $0 < n < \infty$ produzieren Werte relativ dicht am Originalwert mit geringen Abweichungen
- UNS produziert kleinste Werte (19,6% – 65,9% des Originalwertes)

Eval. – Ergebnisse: Clusterkoeffizient

TECHNISCHE UNIVERSITÄT DARMSTADT

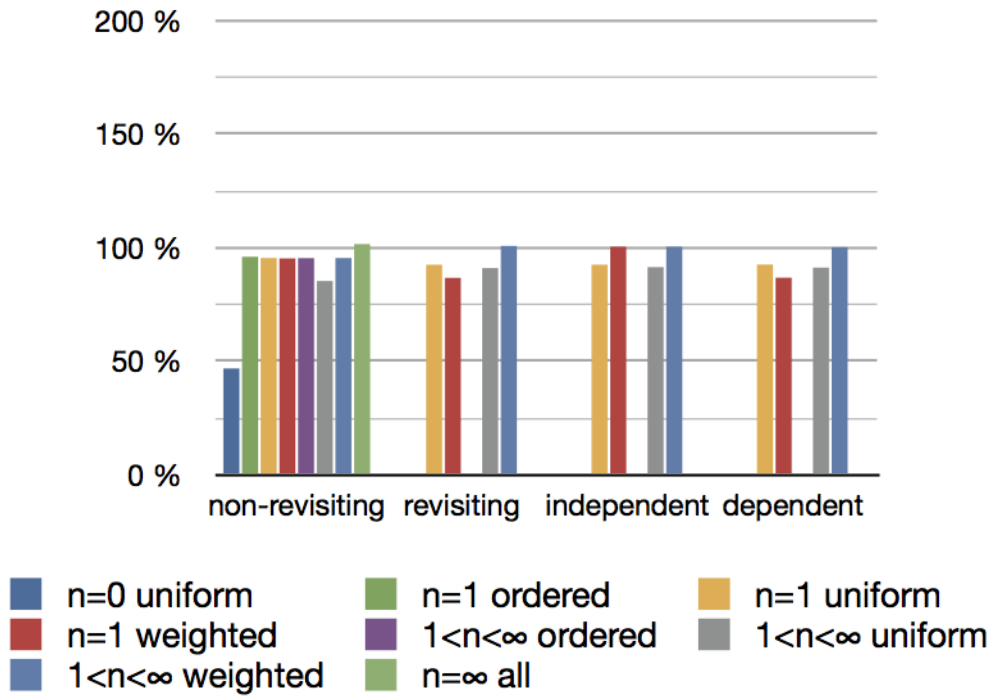
Erdős–Rényi



Eval. – Ergebnisse: Clusterkoeffizient

TECHNISCHE UNIVERSITÄT DARMSTADT

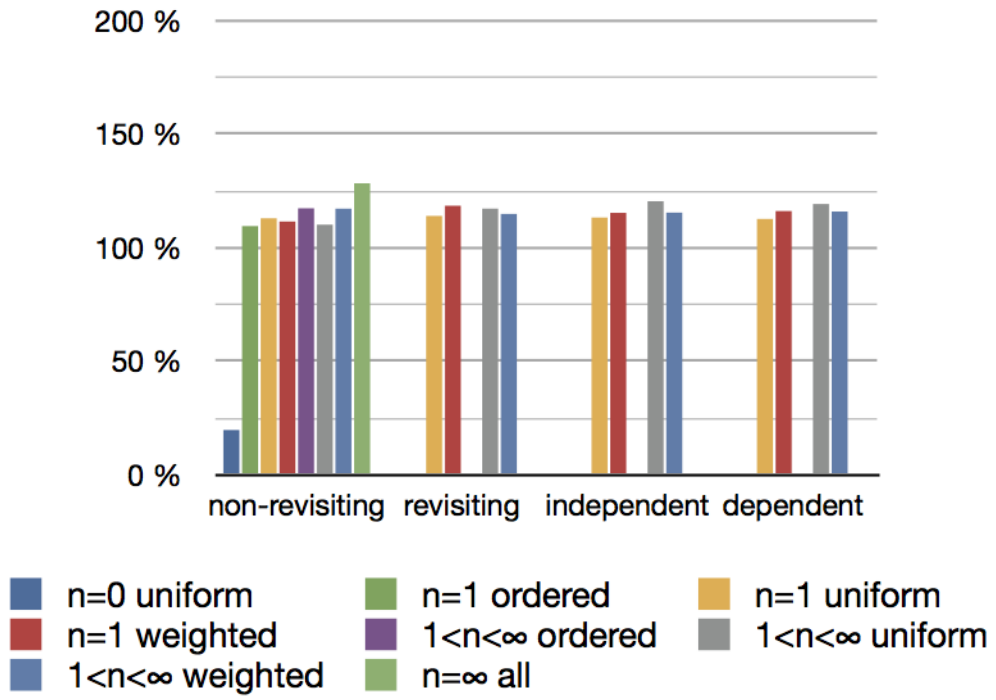
Barabási–Albert



Eval. – Ergebnisse: Clusterkoeffizient

TECHNISCHE UNIVERSITÄT DARMSTADT

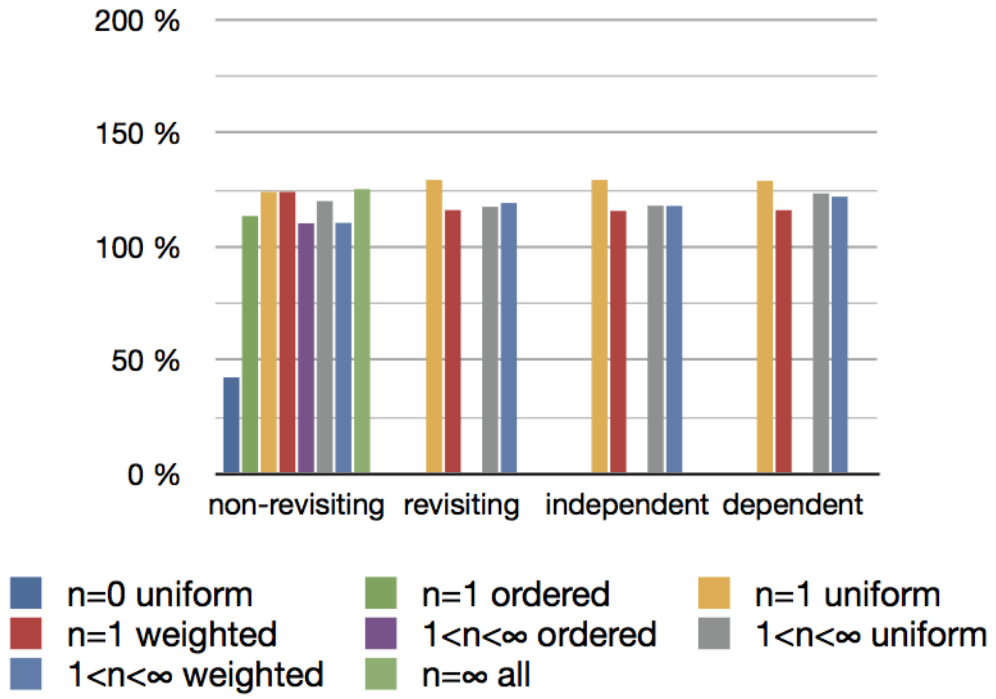
Studentenportal Ilmenau



Eval. – Ergebnisse: Clusterkoeffizient

TECHNISCHE UNIVERSITÄT DARMSTADT

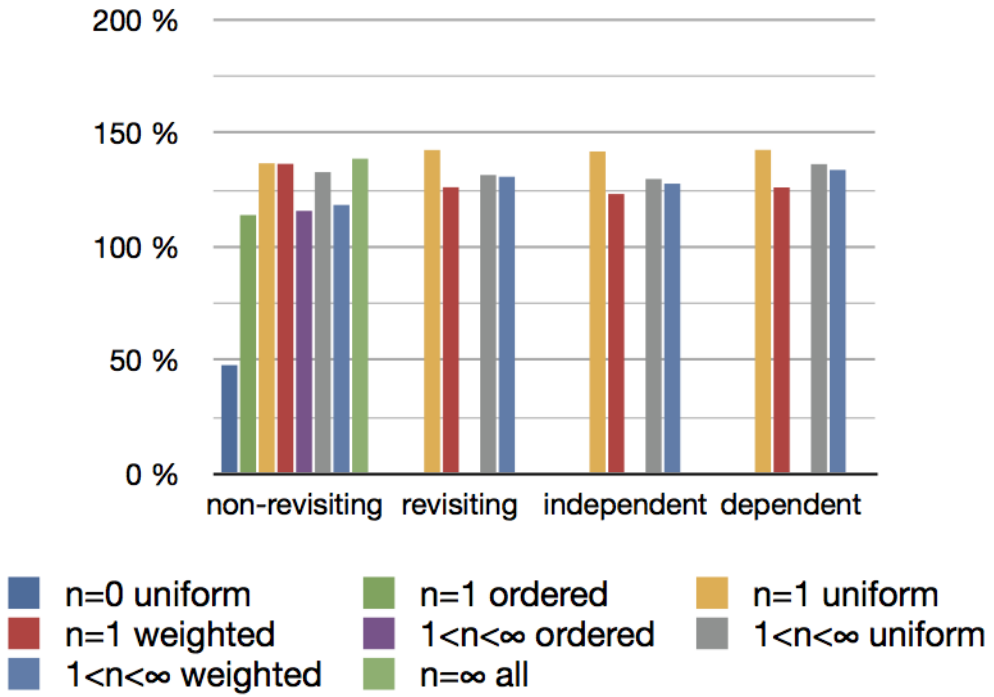
Web of Trust (ungerichtet)



Eval. – Ergebnisse: Clusterkoeffizient

TECHNISCHE UNIVERSITÄT DARMSTADT

Web of Trust (gerichtet)



Zusammenfassung

TECHNISCHE UNIVERSITÄT DARMSTADT

- Subsampling-Algorithmen zweidimensional klassifiziert nach Nachbarwahl und dimensionaler Strategie
- Implementierung
- Evaluation

Fazit

TECHNISCHE UNIVERSITÄT DARMSTADT

- Netzwerkgröße primär durch Nachbarwahl und sekundär durch dimensionale Strategie beeinflusst
- Kombination von Konzepten bewirkt Kombination bzw. Verstärkung von Ergebnissen der Gradverteilung
- DFS-Varianten am originaltreuesten für Durchmesser und Charakteristische Pfadlänge
- Clusterkoeffizient gut bewahrt für $0 < n < \infty$
- $n = 0$ und $n = \infty$ produzieren extreme Ergebnisse

Subsampling of Complex Networks