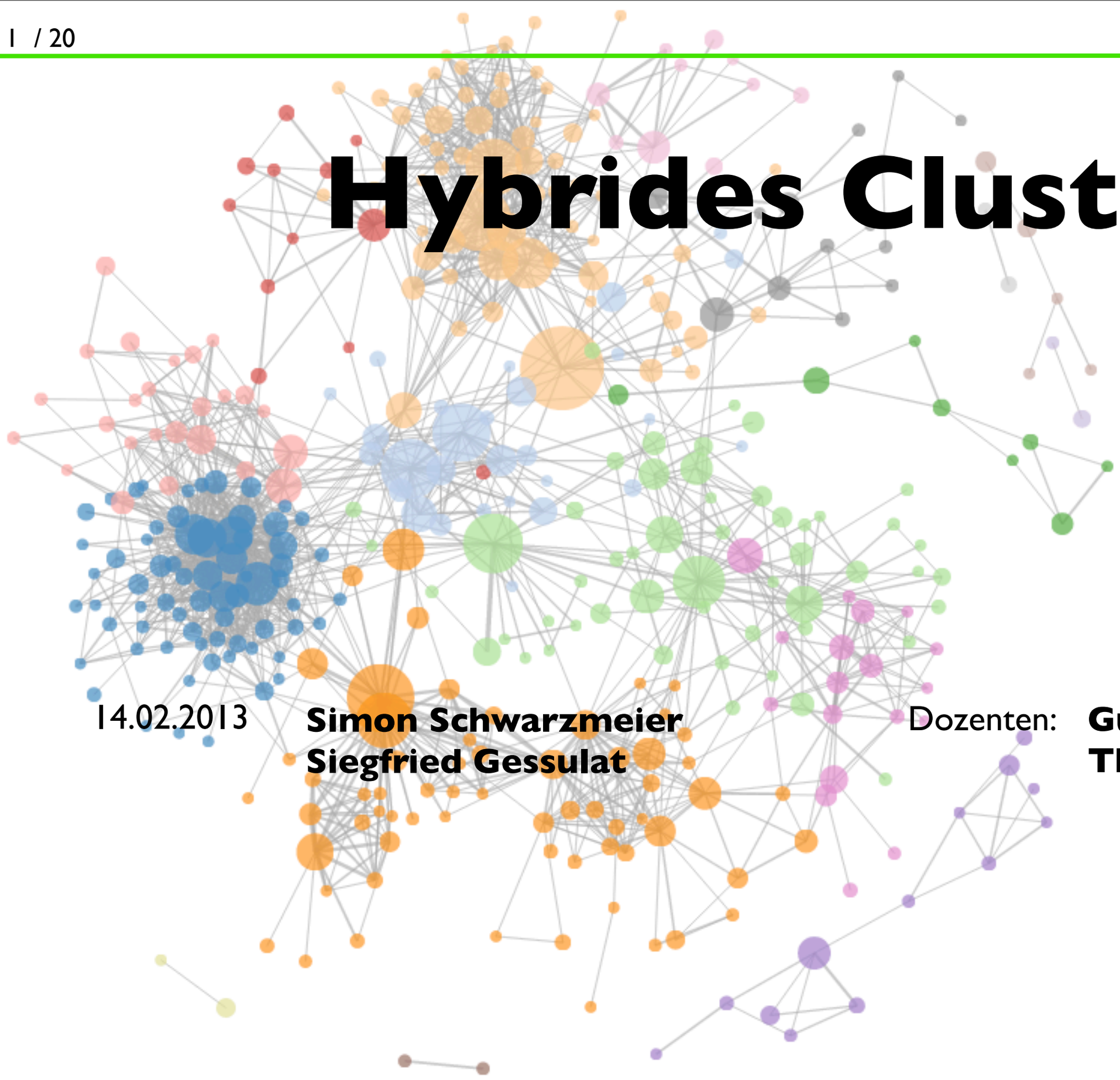
Freie Universität Berlin

Hybrides Clustering



14.02.2013

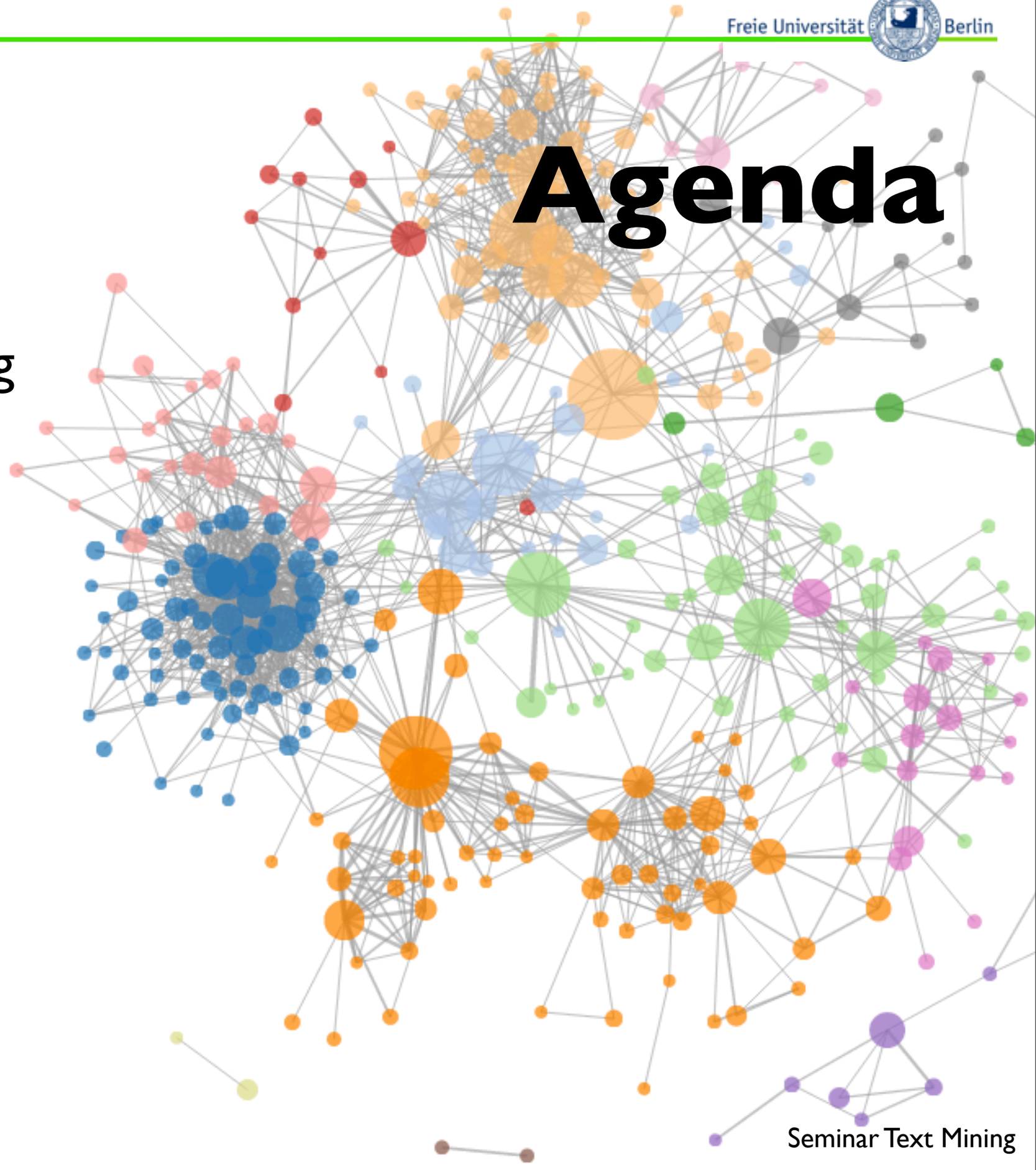
Simon Schwarzmeier
Siegfried Gessulat

Dozenten: **Gunar Maiwald**
Thoralf Klein

Seminar Text Mining

Agenda

1. **Einleitung**
   1. Problemstellung
   2. Recap
2. Methoden
3. Resultate
4. Diskussion



Simon Schwarzmeier
Siegfried Gessulat

Seminar Text Mining

Problemstellung

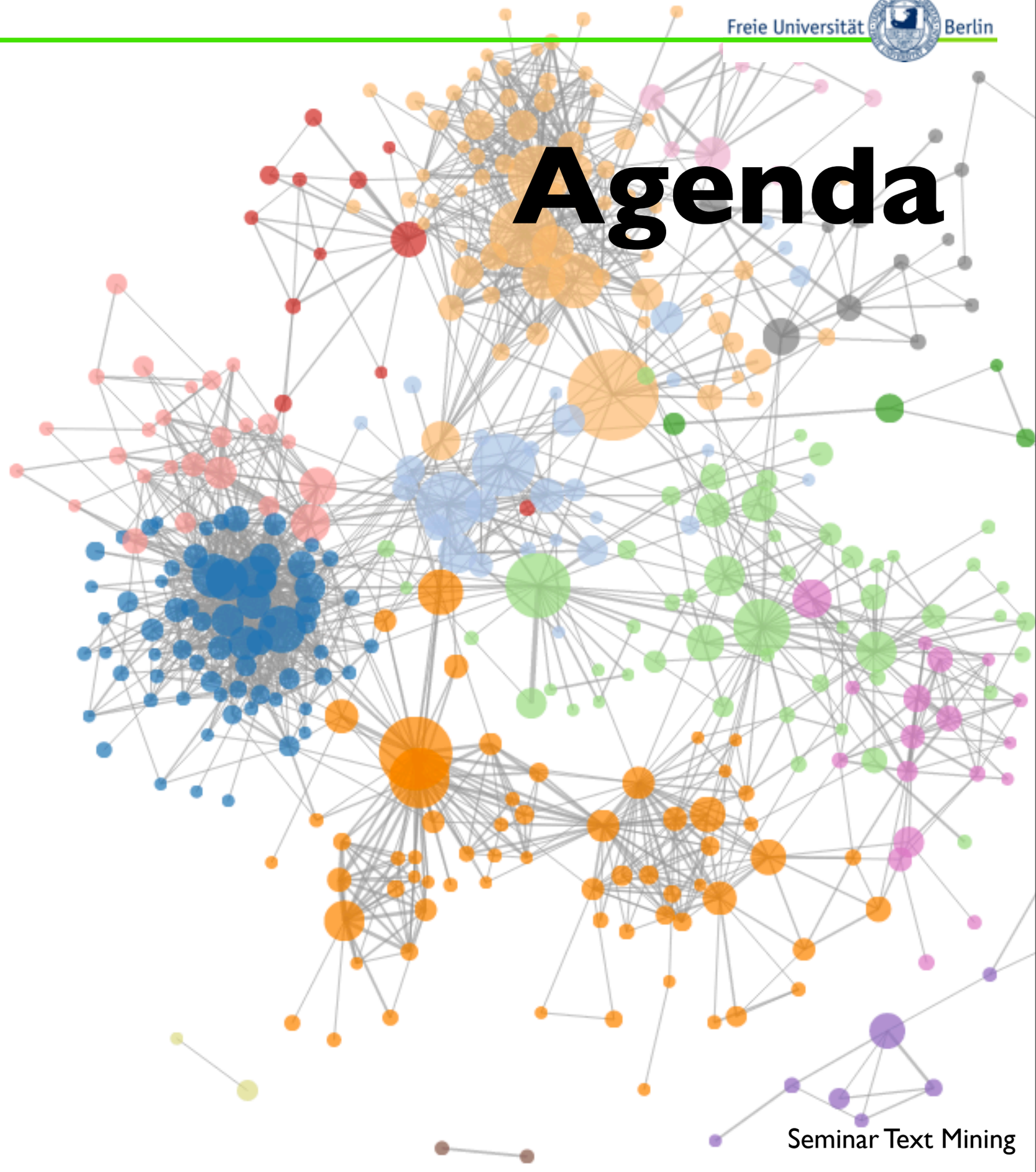
- Clustering von Mathematischen Artikeln
- Distanzmaß ist eine Mischung aus zwei verschiedenen Maßen (Cosinus und Tanimoto)
- Clustering mit k-means dialekt: k-medoids
- Initialisierung von k-medoids mit einer neuen Methode (h-cores)

Recap

- Rogers-Tanimoto distance: Abstandsmaß basierend auf **Zitationsgraph**
- Cosine distance (VSM): räumliches Abstandsmaß basierend auf **abstracts**
- Fisher's Inverse Chi-Square Methode: gewichtete **Mischung** zweier Distanzen. Parameter: lambda (Gewichtung)
- H-Cores: Methode um initiale **Clusterzentren** zu finden. Parameter: r (Schwellwert)

Agenda

1. Einleitung
2. **Methoden**
   1. Preprocessing
   2. Clustering
   3. Tools
3. Resultate
4. Diskussion



Simon Schwarzmeier
Siegfried Gessulat

2. Methoden

Preprocessing

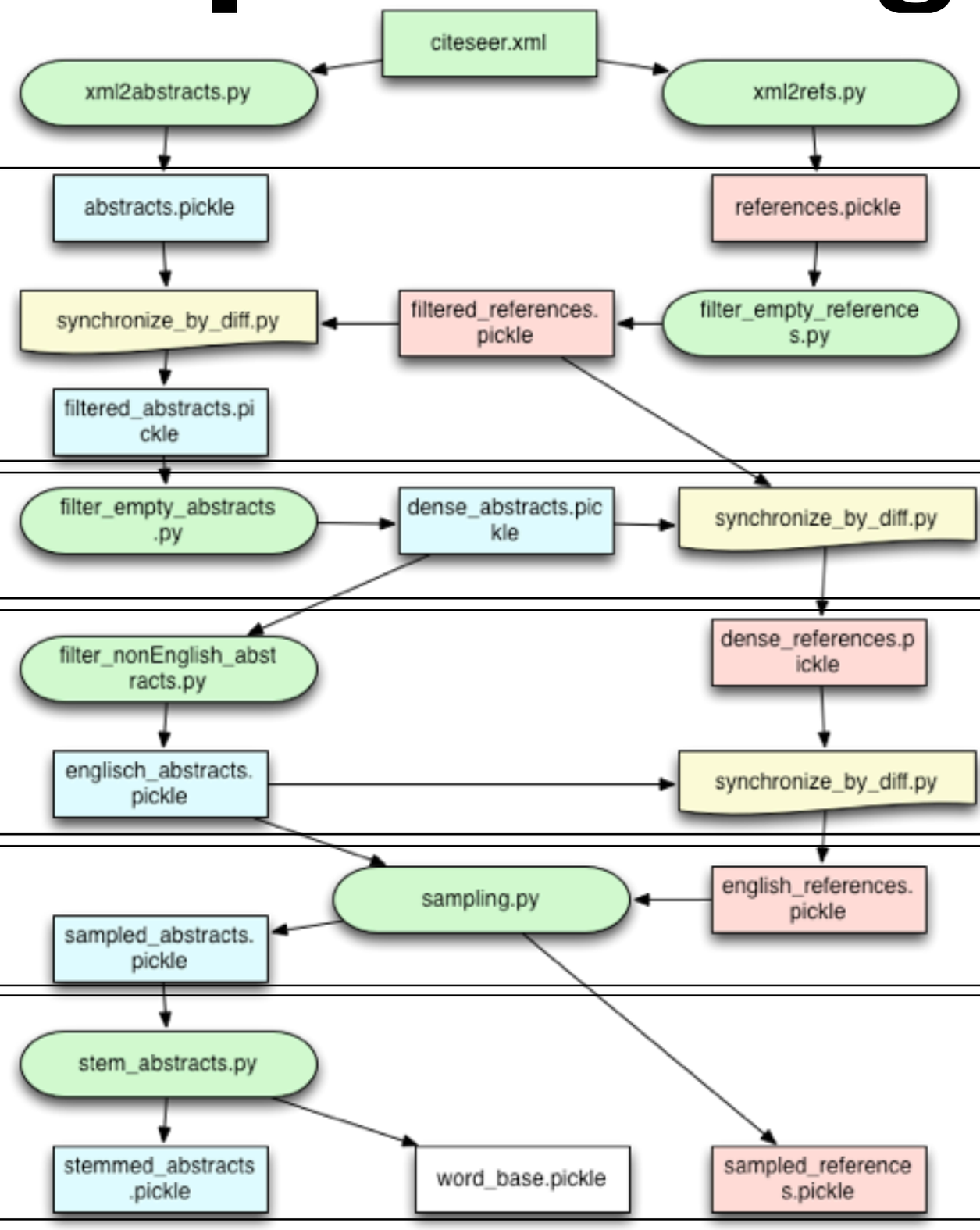
leere Refernzen filtern
(tanimoto distance)

leere Abstracts filtern
(cosinus distance)

nicht-englische Abstracts filtern

Daten reduzieren durch Sampling
random und **meiste Referenzen**

Stemming und Generierung der
word_base für das VSM



2. Methoden

Clustering

- Sampling Methoden:
  - random: 10k zufällige Dokumente
  - most refs: 10k am häufigsten referenzierten Dokumente
- Parameter
  - lambda: gibt die Mischung der Beiden Maße an
  - r: Schwellwert für die h-cores und damit indirekt das k für k-medoids

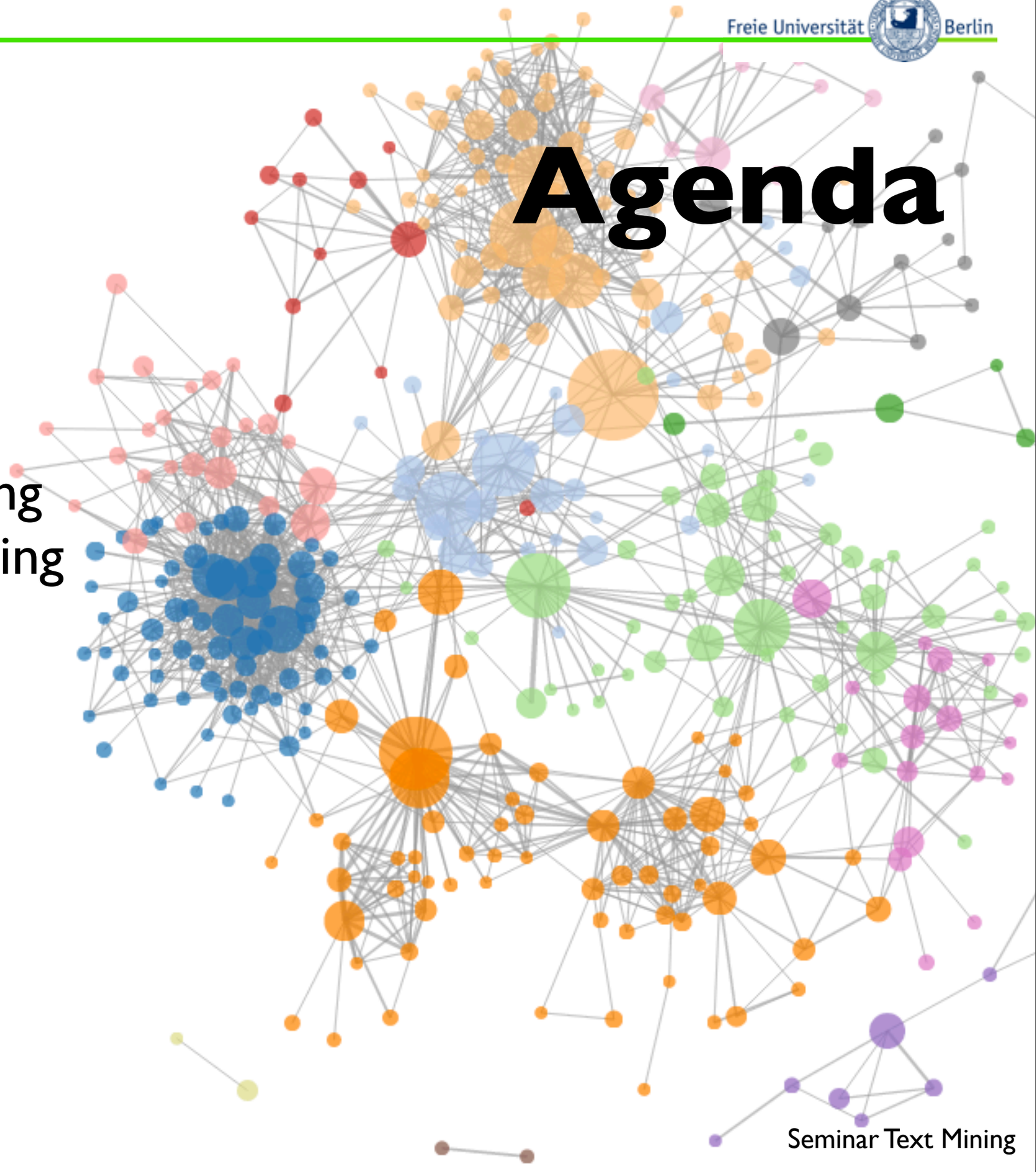
2. Methoden

Tools

- **XML Parsing:** xml.sax
- **Stemming:** stemming.porter2 stem
- **Tokenizing:** nltk.tokenize: wordpunct_tokenize
- **Spracherkennung:** guess-language
- **Distanzen:** scipy.spatial distances
- **Clustering:** C Clustering Library (Univ. Tokyo) mit Python Bindings
- **Animationen:** ProgressFish ><((((('>

Agenda

1. Einleitung
2. Methoden
3. **Resultate**
   1. random Sampling
   2. most ref. Sampling
   3. Clusterqualität
4. Diskussion



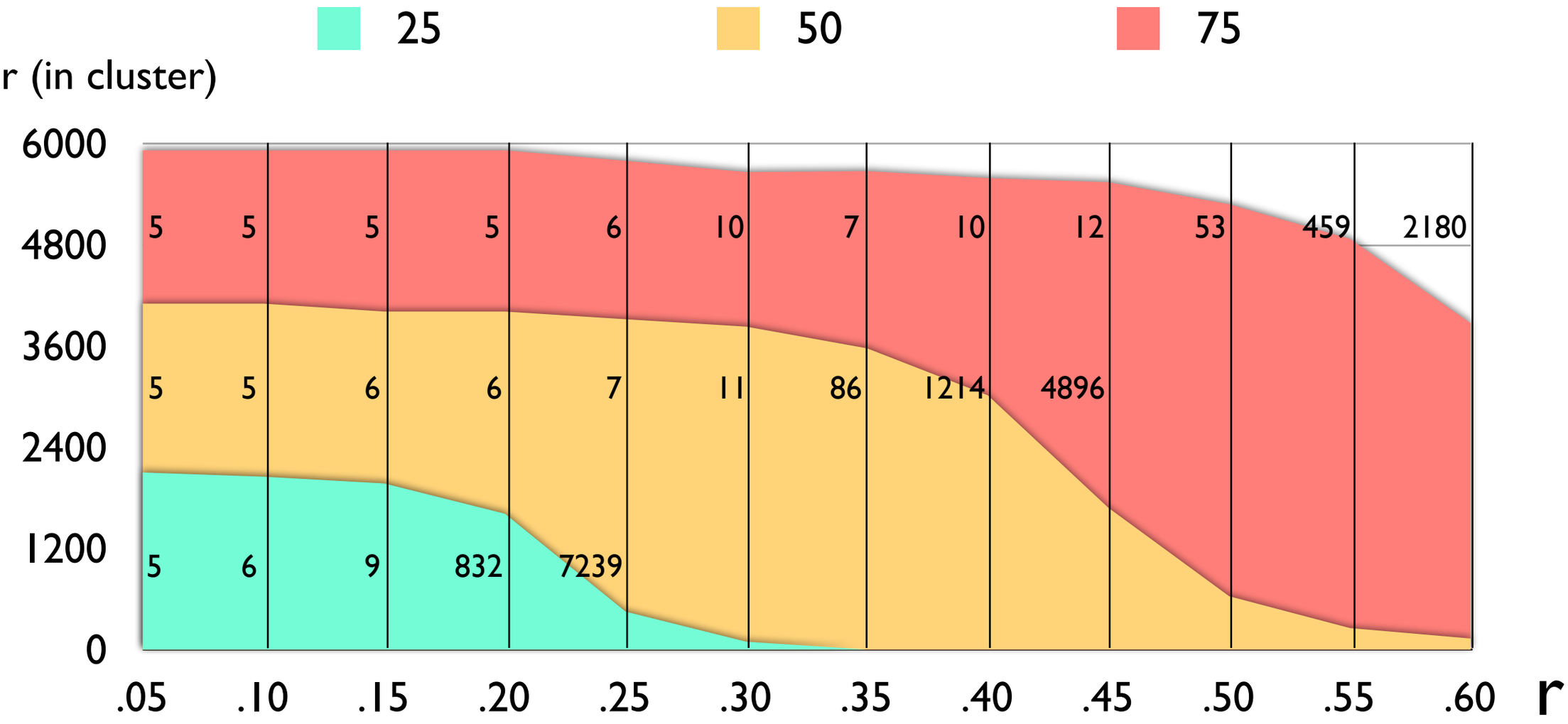
Simon Schwarzmeier
Siegfried Gessulat

Seminar Text Mining

Freie Universität Berlin

3. Resultate

random references

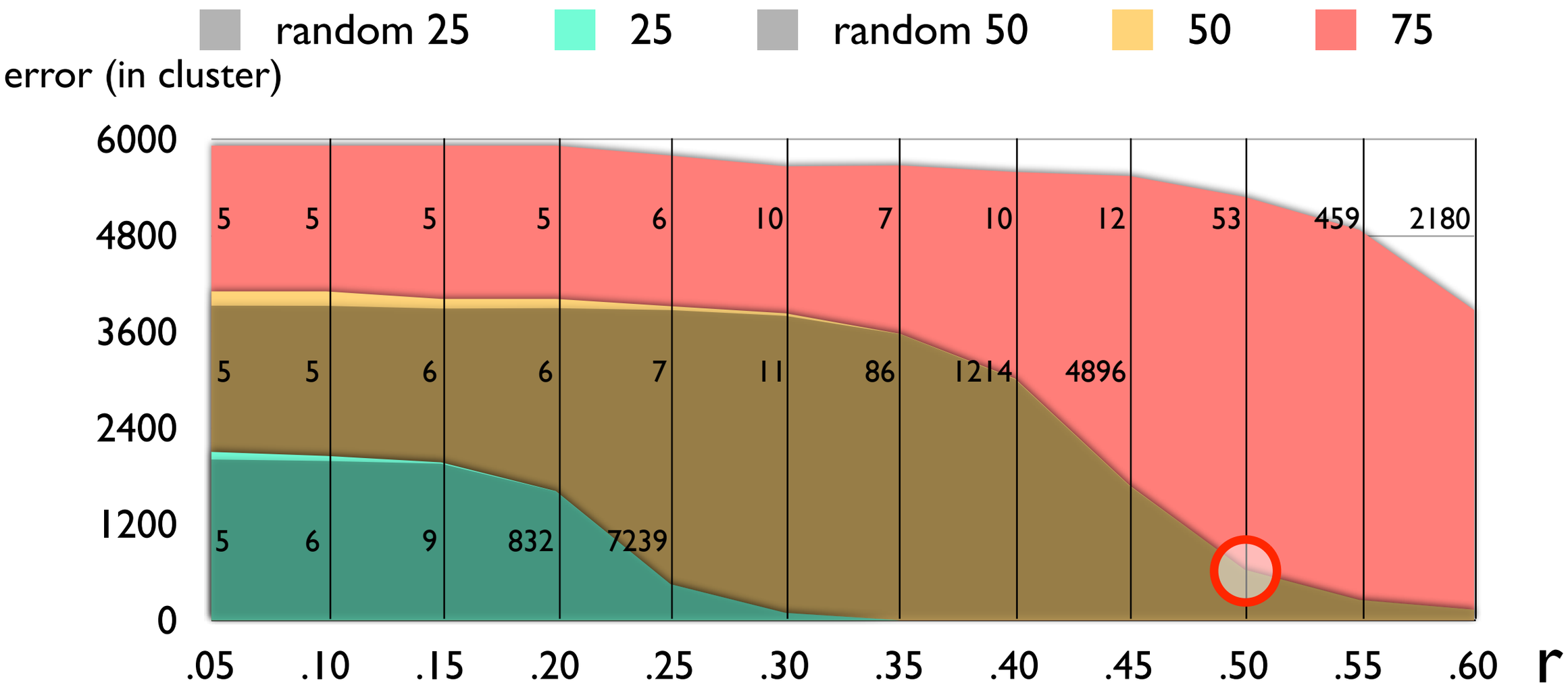


Simon Schwarzmeier
Siegfried Gessulat

Freie Universität Berlin

3. Resultate

random references

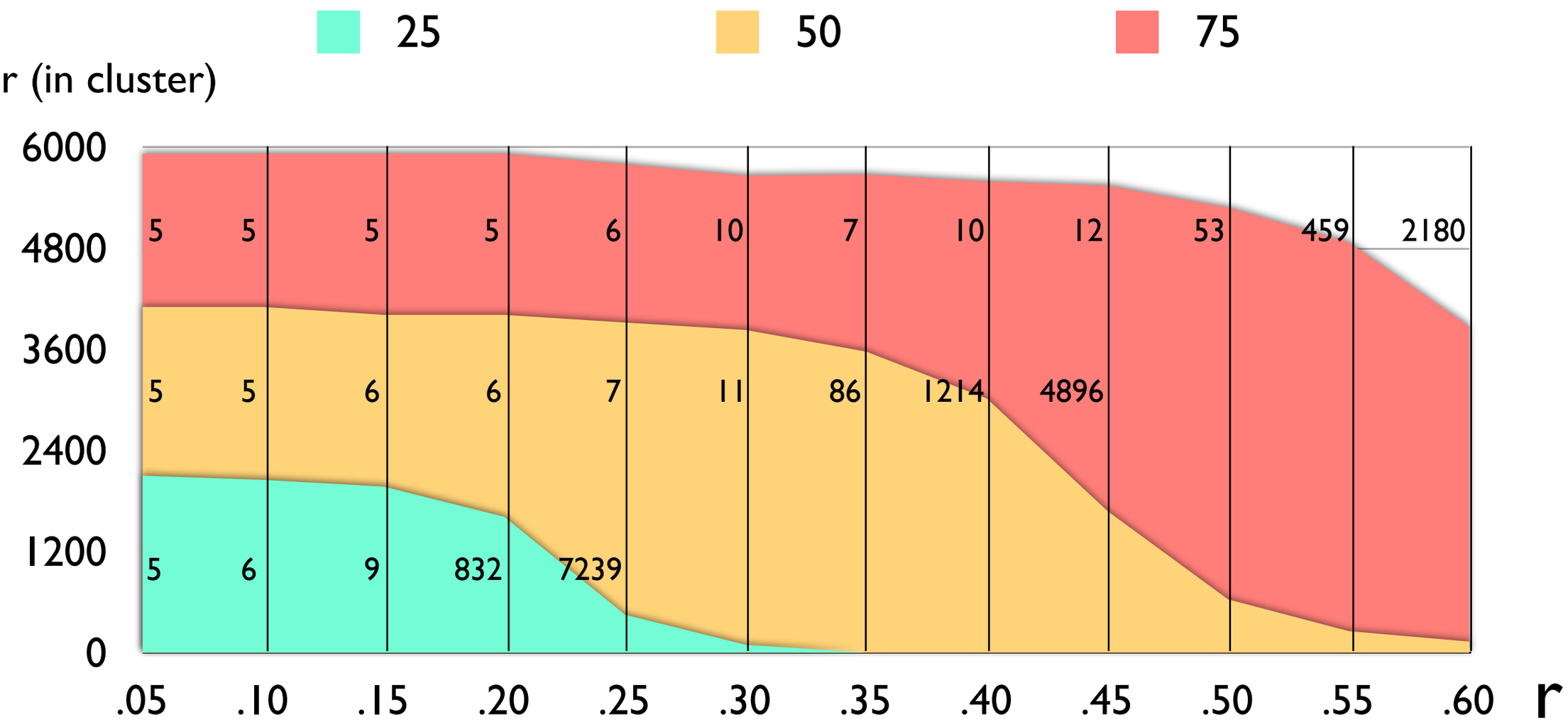


Simon Schwarzmeier
Siegfried Gessulat

Freie Universität Berlin

3. Resultate

random references

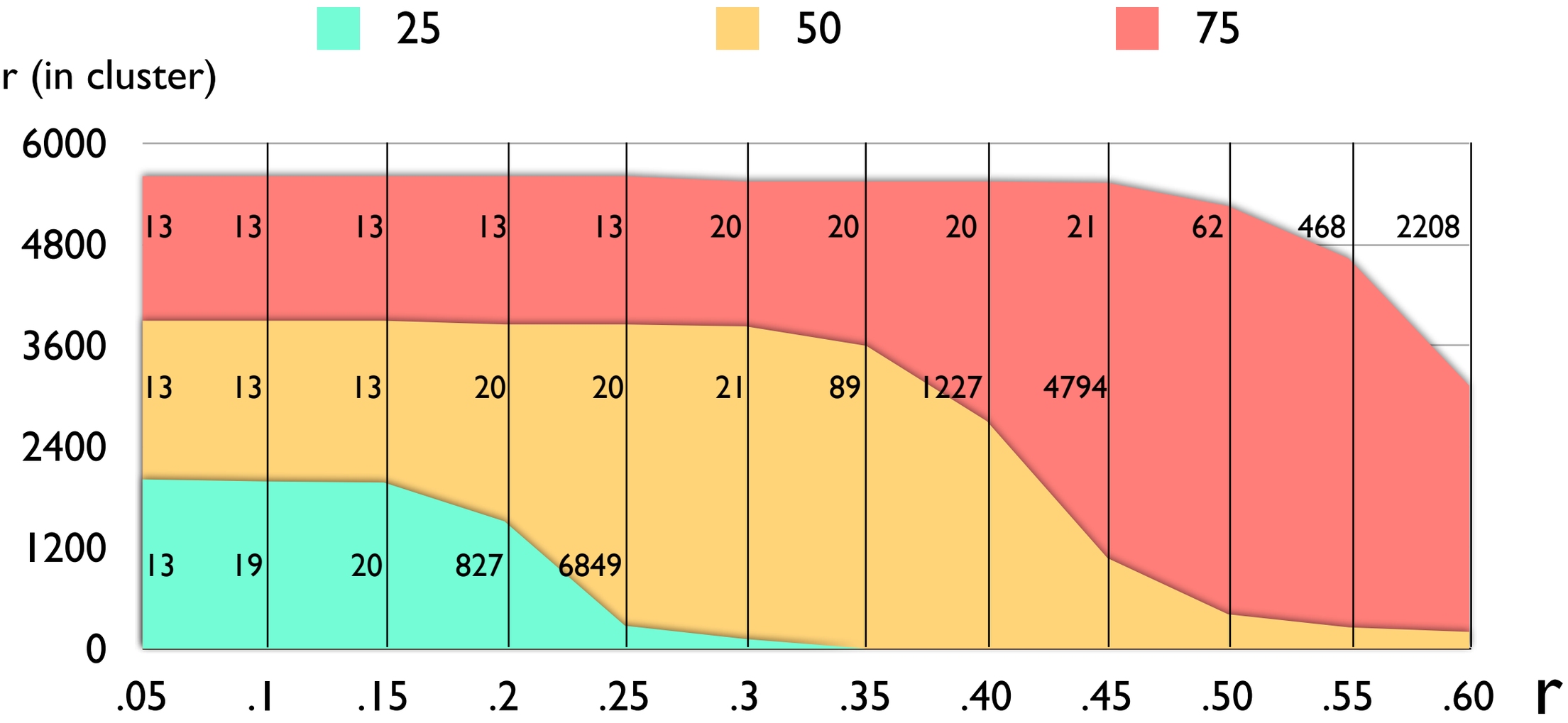


Simon Schwarzmeier
Siegfried Gessulat

Freie Universität Berlin

3. Resultate

most references

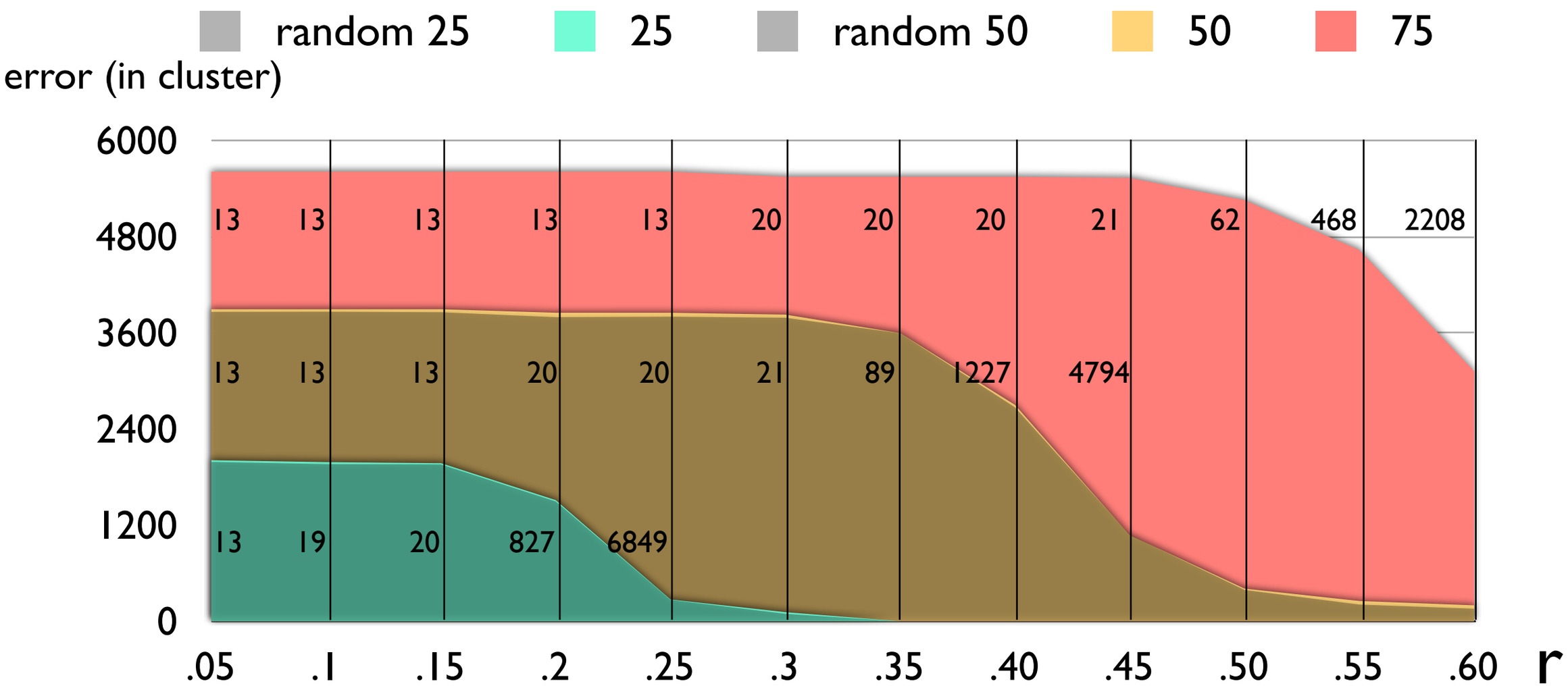


Simon Schwarzmeier
Siegfried Gessulat

Freie Universität Berlin

3. Resultate

most references



Simon Schwarzmeier
Siegfried Gessulat

6157

paper	397
system	341
1	308
present	304
base	281
algorithm	275
problem	264
perform	261
model	255
result	240

6342

paper	444
present	358
system	337
base	316
1	312
perform	294
result	291
algorithm	273
model	270
show	260

4702

paper	338
system	295
present	274
1	247
base	220
algorithm	213
problem	209
perform	198
applic	194
data	192

5704

paper	236
system	193
present	193
1	162
base	161
problem	155
model	154
algorithm	151
data	139
result	138

3494

paper	258
system	220
present	195
1	170
base	168
result	156
problem	154
provid	153
algorithm	144
perform	144

3495

paper	252
system	208
base	176
1	174
present	169
algorithm	160
perform	156
result	155
problem	141
model	141

6723

paper	293
present	247
base	246
system	245
1	238
perform	208
data	206
model	198
result	192
problem	188

5952

paper	231
present	196
system	191
1	178
base	175
algorithm	162
applic	157
provid	145
result	144
model	141

5659

paper	396
system	304
present	276
1	275
base	274
data	258
perform	236
algorithm	236
model	233
result	230

4556

paper	228
present	161
system	158
base	148
1	140
perform	134
algorithm	132
data	132
result	127
problem	126

5643

paper	330
system	310
present	254
base	239
1	224
algorithm	223
model	209
perform	199
provid	192
problem	190

6175

paper	301
system	231
present	220
1	217
base	206
algorithm	205
model	188
result	186
problem	178
comput	166

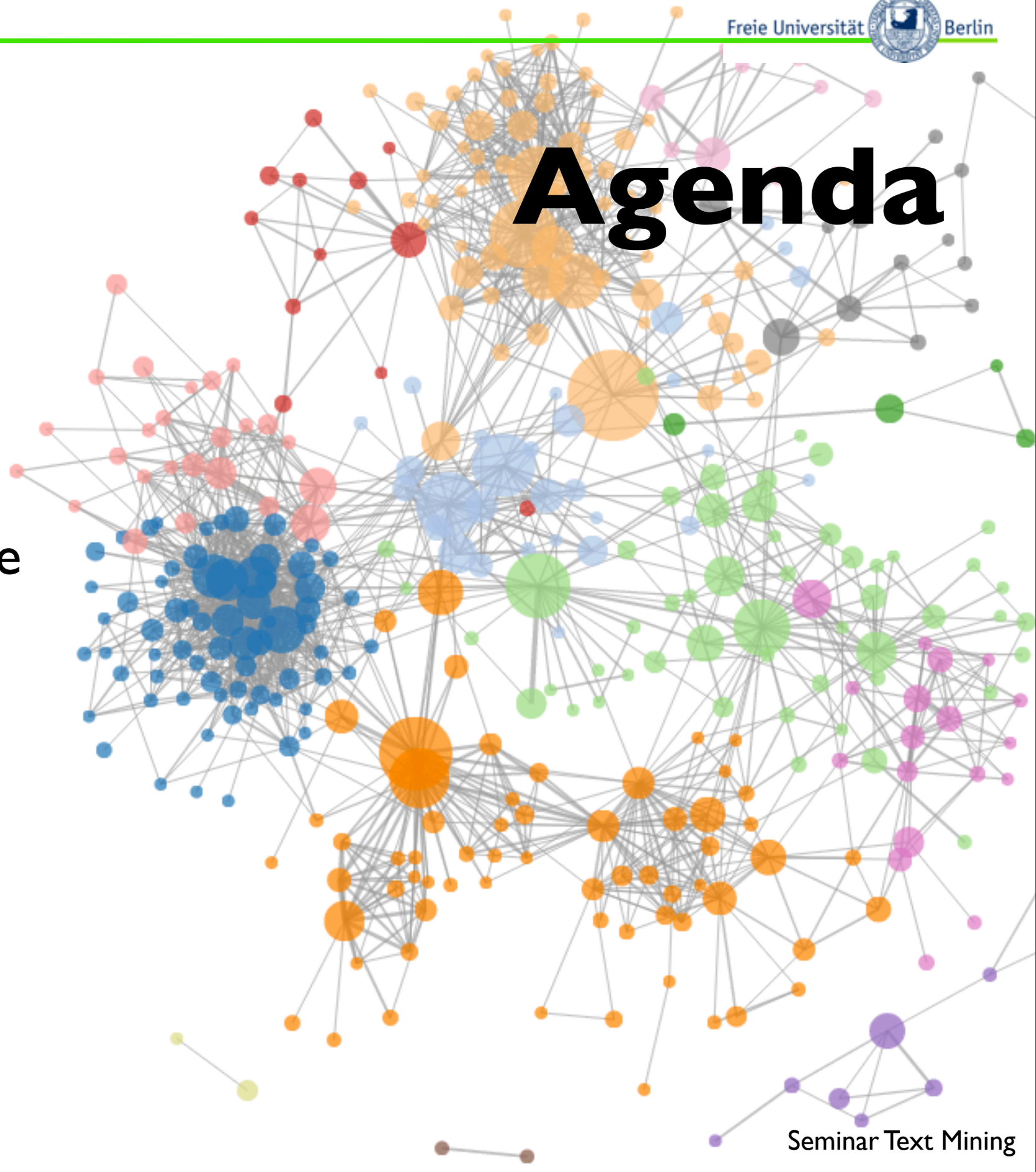
6233

paper	358
present	285
system	283
1	262
result	248
base	244
algorithm	239
perform	233
data	218
comput	214

Simon Schwarzmeier
Siegfried Gessulat

Agenda

1. Einleitung
2. Methoden
3. Resultate
4. **Diskussion**
   1. Positive Aspekte
   2. Schwächen
   3. Interpretation
   4. Fragen

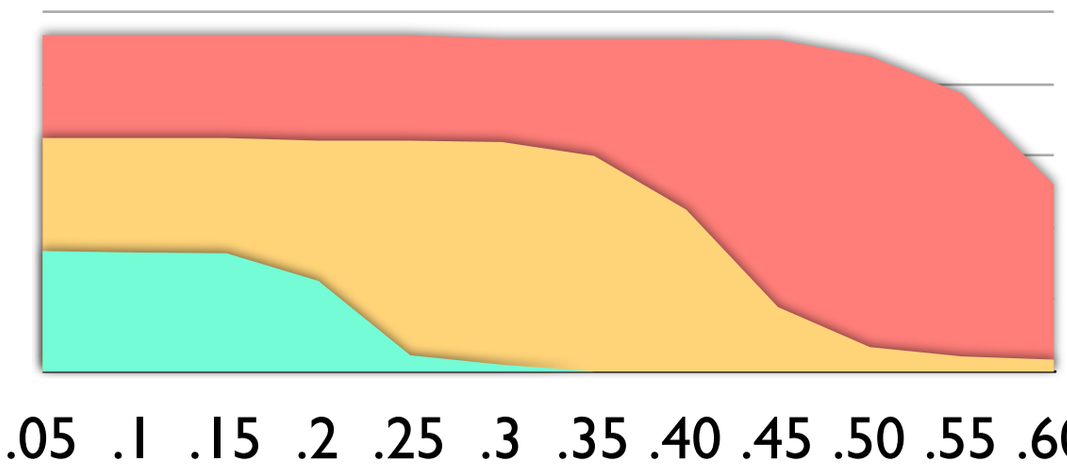


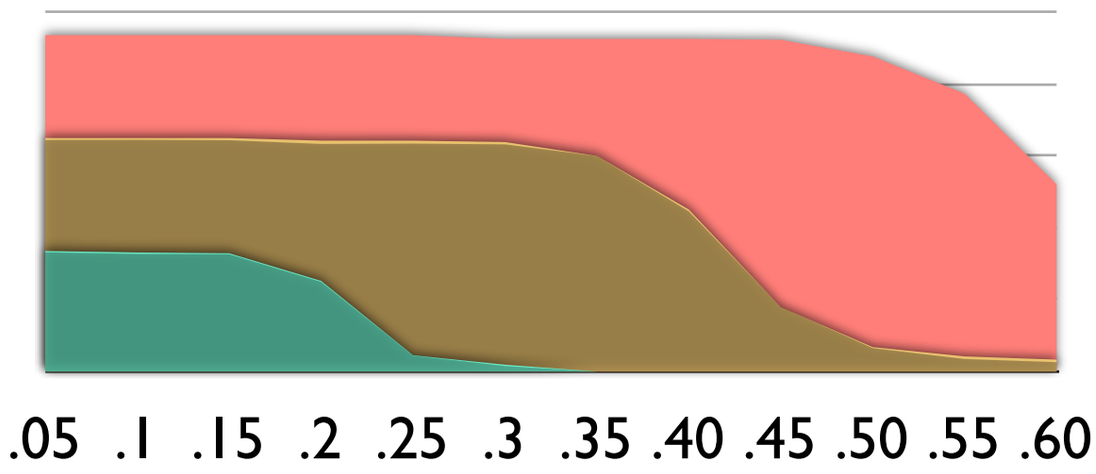
Simon Schwarzmeier
Siegfried Gessulat

4. Diskussion

Positive Aspekte

- wir nehmen an: grau ist das Optimum.
- wir erreichen fast den gleichen Wert mit nur einem Durchlauf (deterministisch!)
- most refs Datensatz funktioniert besser.





4. Diskussion

Schwächen

- wir benutzen nur ~1% der Daten
- nur zwei sampling Methoden und nur jeweils ein sample set.
- VSM: wir benutzen kein TF.IDF
- Error nur innerhalb des Clusters, nicht in bezug zu anderen Clustern

4. Diskussion

Intepretation

- Güte der Cluster ist schwer beschreibbar, weil es eine Verteilung über alle Wörter ist

- Initialisieriung durch h-cores: Lohnt sich der Aufwand? Nein, clustern ist nicht bottle-kneck (bei uns)
- Wenn man h-cores nicht braucht, braucht man dann k-medoids? LDA?

Freie Universität Berlin

4. Diskussion

Fragen!

WTF?!

Simon Schwarzmeier
Siegfried Gessulat