Influence Attribution in Social Networks

Panagiotis Papapetrou

Aalto University, Finland

P. Papapetrou, Aris Gionis, and Heikki Mannila, "A Shapley value Approach for Influence Attribution" *ECML-PKDD 2011*

Influential individuals

- People always intrigued by characterizing influential ideas, books, scientists, politicians, etc.
- Main question: who is influential?
- Examples
  - Who initiates the most influential “tweets”?
  - Who are the most influential scientists?
  - Which actors influence a movie rating the most?

Our setting

- Individuals accomplish tasks in a collaborative manner.
- **Influence attribution:** each individual is assigned a score based on his/her performance.

Example: author-publication

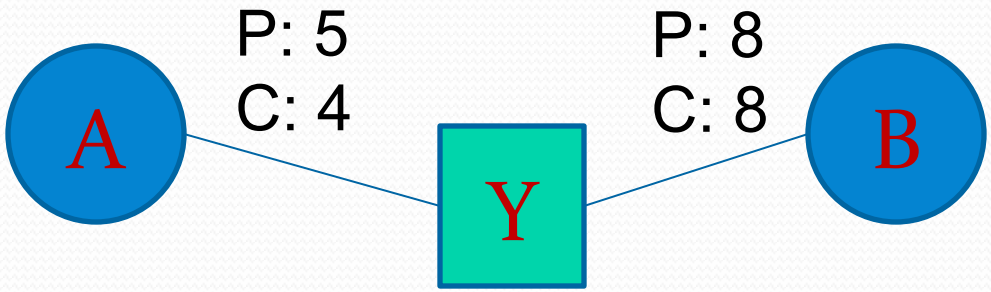
- Individual => author.
- Task => publication.
- Impact score:
  - CC: Citation count of the publication.
  - PR: PageRank score of the publication.

Example: author-publication

- Two researchers A and B.
- Question: who is more influential?

Example: author-publication

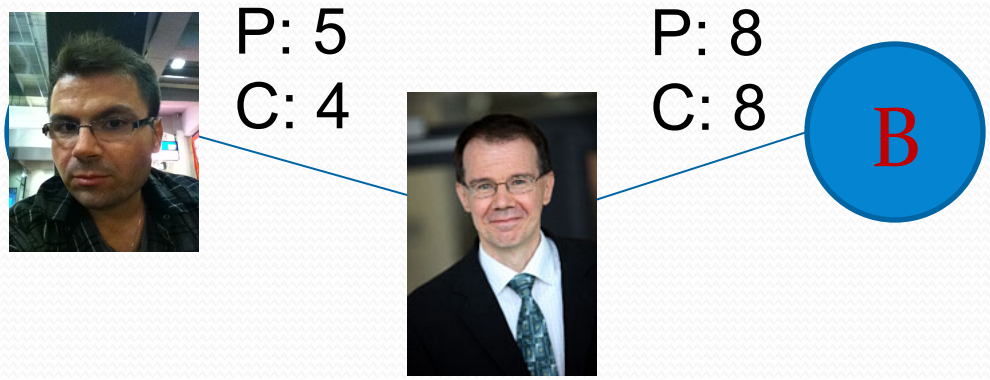
- One common collaborator: Y.



P: number of papers
C: number of citations per paper

Example: author-publication

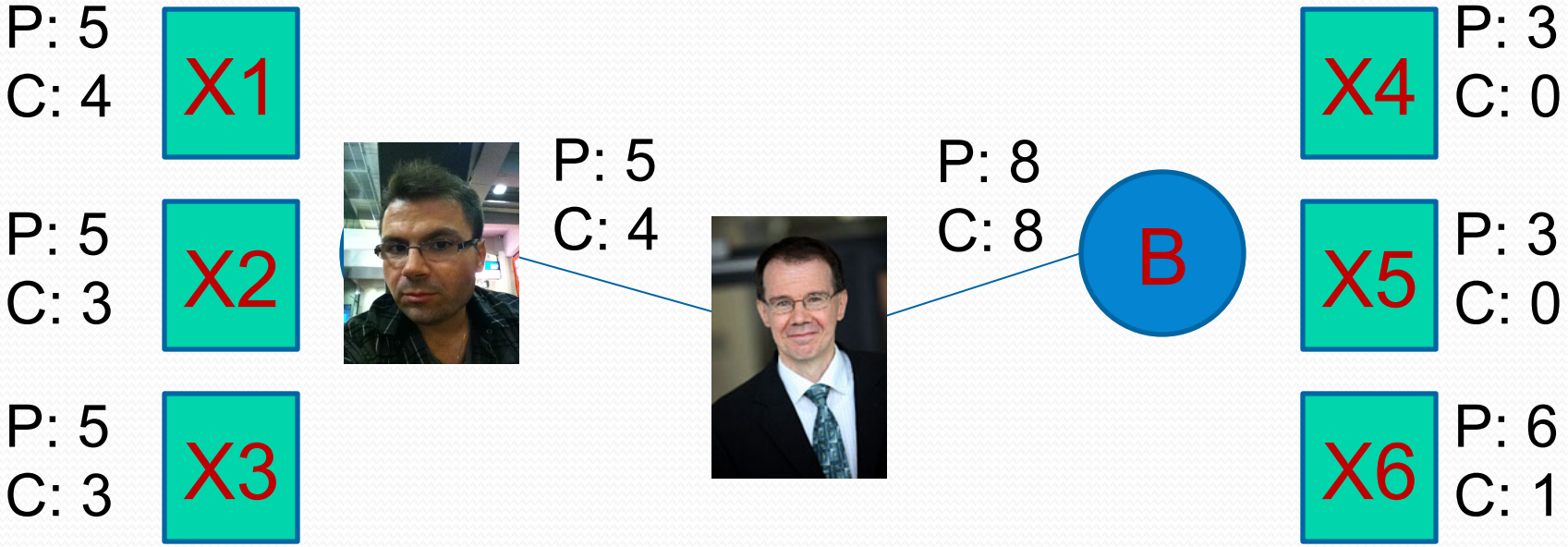
- One common collaborator: Y.



P: number of papers
C: number of citations per paper

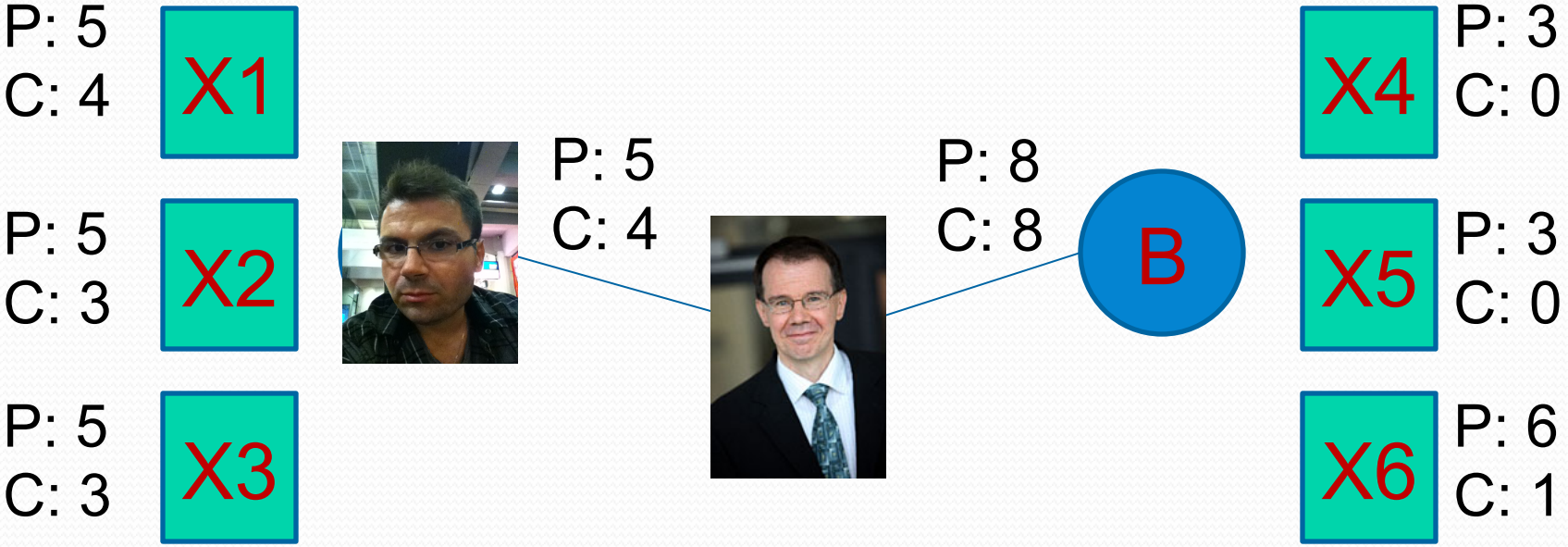
Example: author-publication

- Three additional collaborators for A and B.



Example: author-publication

- Three additional collaborators for A and B.



Researcher	Papers	Citations	H-index
A	20	70	4
B	20	70	8

Example: author-publication

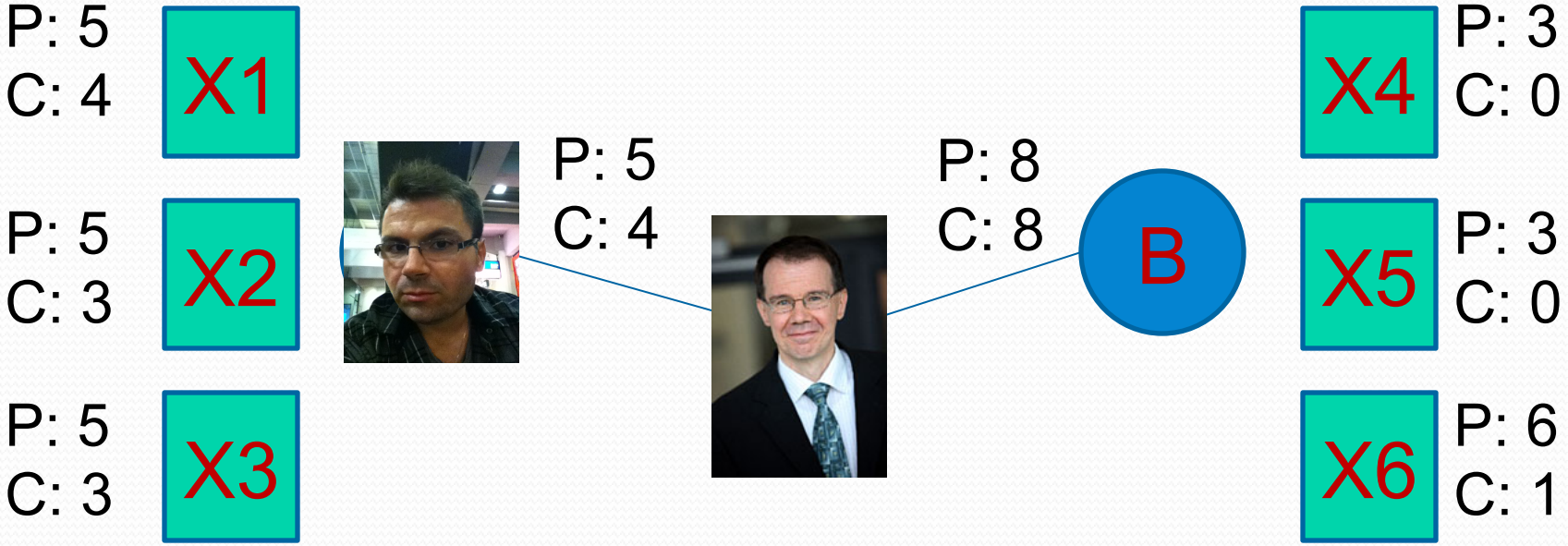
- Three additional collaborators for A and B.

H-Index: a scientist's H-index is h, if h of his/her publications have at least h citations and the rest of his/her publications have at most h citations each.

Researcher	Papers	Citations	H-index
A	20	70	4
B	20	70	8

Example: author-publication

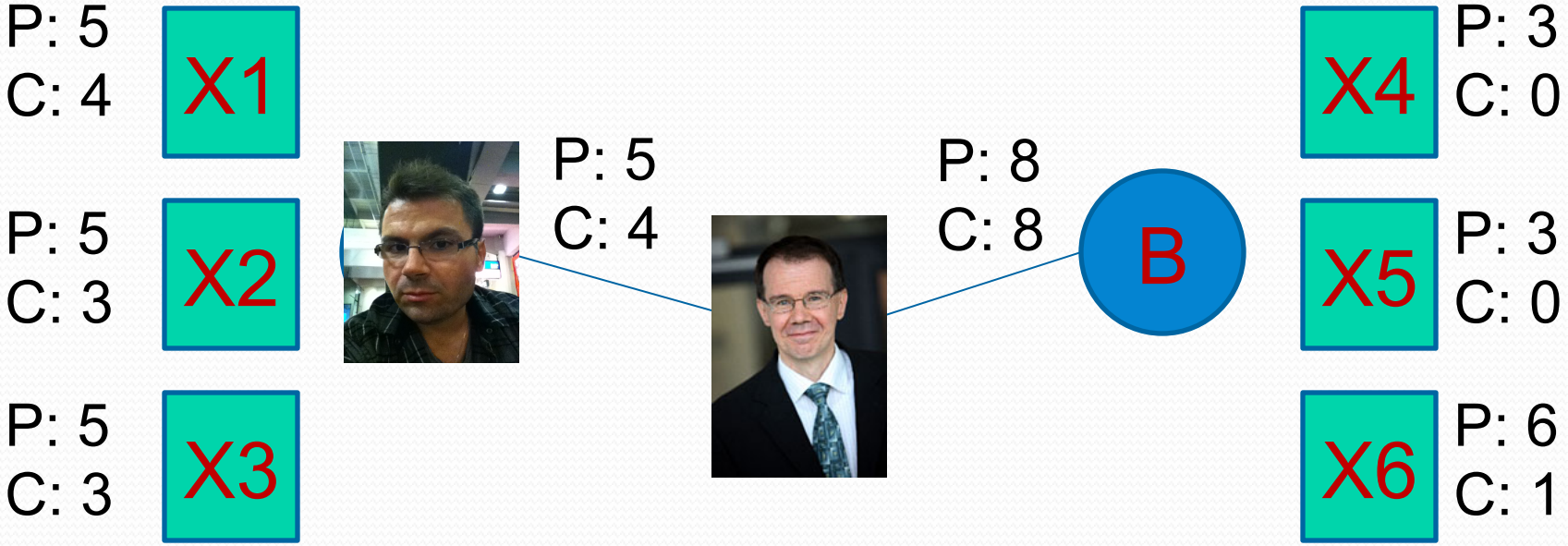
- Three additional collaborators for A and B.



Researcher	Papers	Citations	H-index
A	20	70	4
B	20	70	8

Example: author-publication

- Three additional collaborators for A and B.



- But is B indeed that influential?
- Or is B just being favored due to the fame of Y?

Example: author-publication

- Drop Y out of the picture.

- The performance of A remains quite high.
- The performance of B is weakened a lot.

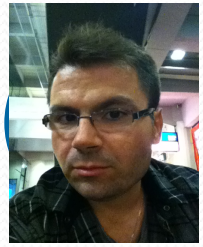
Example: author-publication

- Drop Y out of the picture.

P: 5
C: 4
X1

P: 5
C: 3
X2

P: 5
C: 3
X3



B

X4
P: 3
C: 0

X5
P: 3
C: 0

X6
P: 6
C: 1

Researcher	Papers	Citations	H-index
A	15	50	4
B	12	6	1

Our Approach

- For each individual compute:
  - what difference does an individual make to the coalition if dropped from it.
- Individuals who form many strong coalitions are favored against those who form weaker coalitions.

Shapley Value

- Set of individuals V, set of tasks T, and task impact scores I.
- Gain function $v(\mathcal{S})$
  - gain achieved by the cooperation of the individuals in S.
- Shapley value: the sum of all marginal gains contributed by each individual to a coalition.

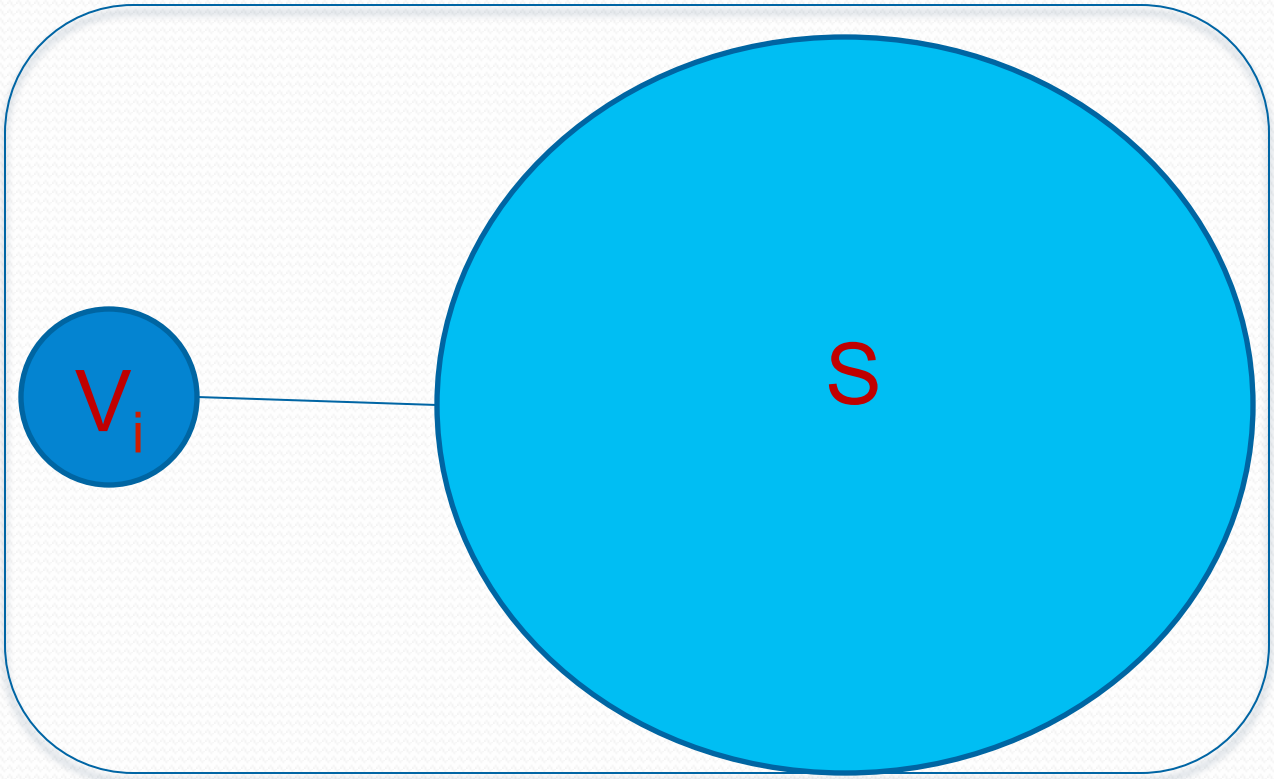
$$\phi_i(v) = \sum_{\mathcal{S} \subseteq \mathcal{V}} \frac{|\mathcal{S}|!(|\mathcal{V}| - |\mathcal{S}| - 1)!}{|\mathcal{V}|!}(v(\mathcal{S} \cup \{V_i\}) - v(\mathcal{S})).$$

Our Approach

- Not all coalitions may be available or defined.
- We compute the marginal gains by averaging only over **coalitions for which impact scores are available**.
- For the author-publication case: iterate over all papers.
- We approximate the rest.

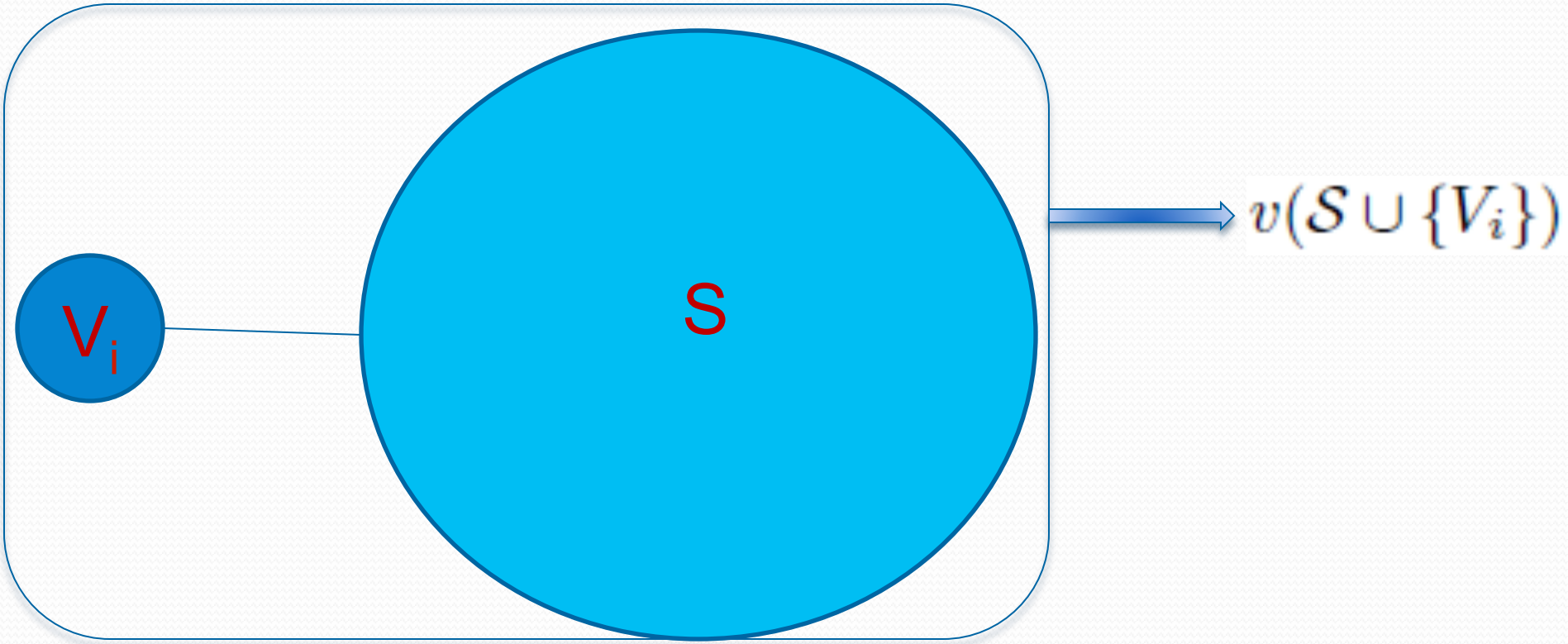
Iterative method

- We choose to take into account all cases for which $S \cup \{V_i\}$ is available.



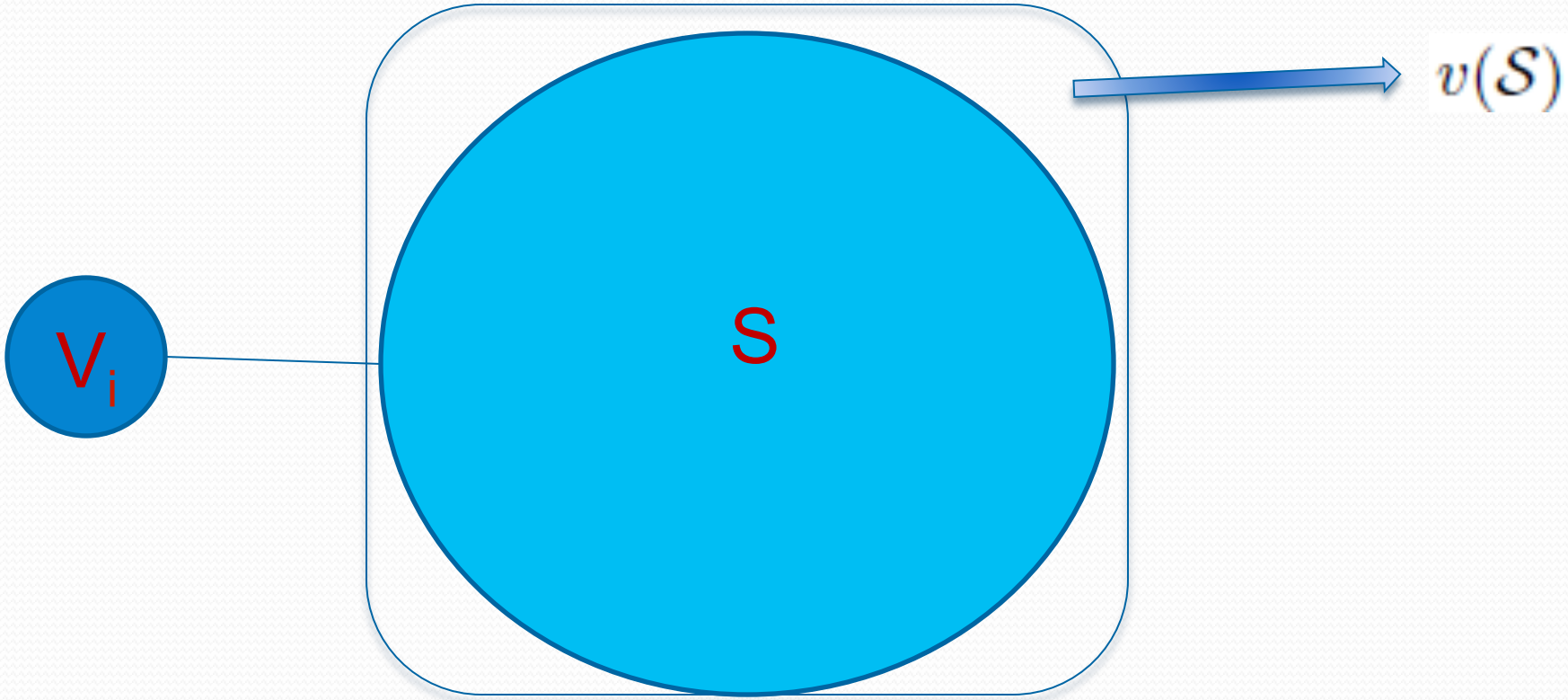
Iterative method

- We choose to take into account all cases for which $\mathcal{S} \cup \{V_i\}$ is available.



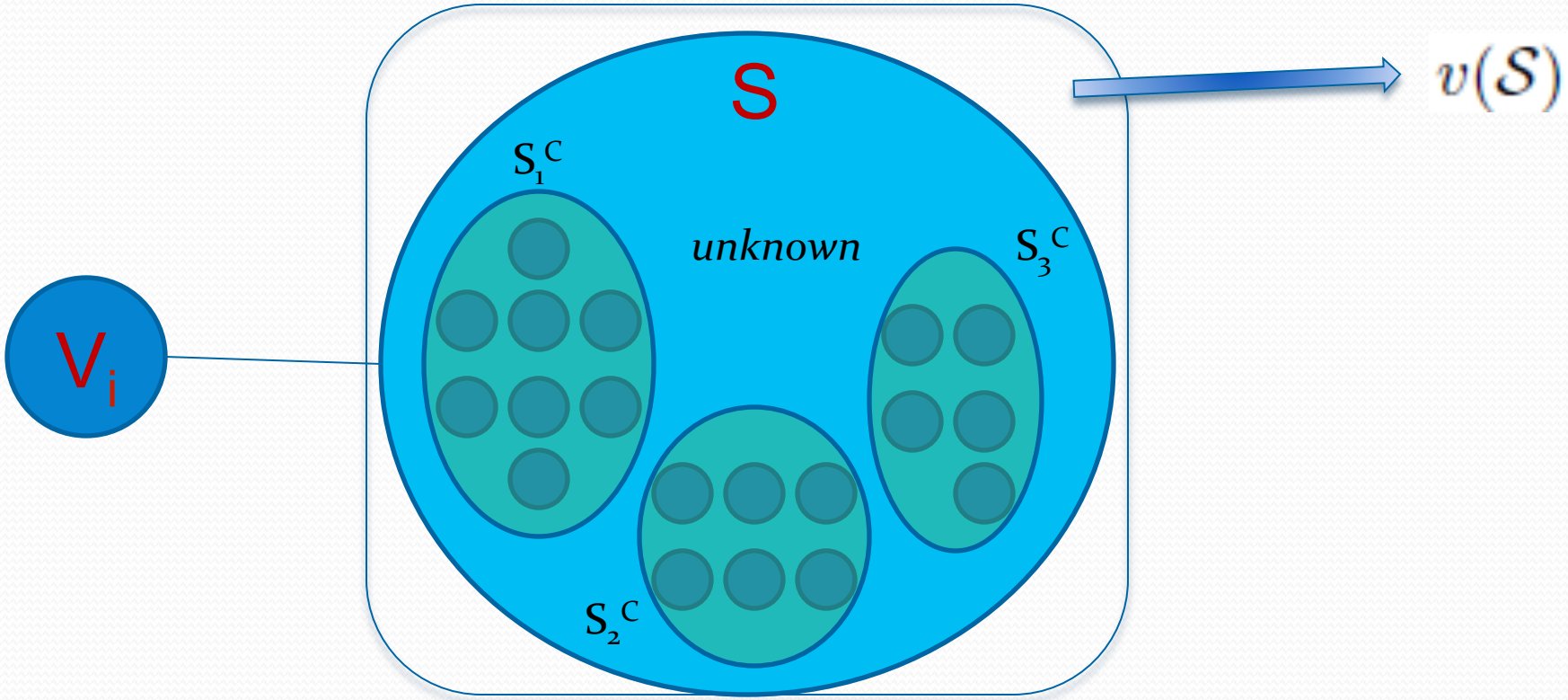
Iterative method

- Then compute the gain of S.



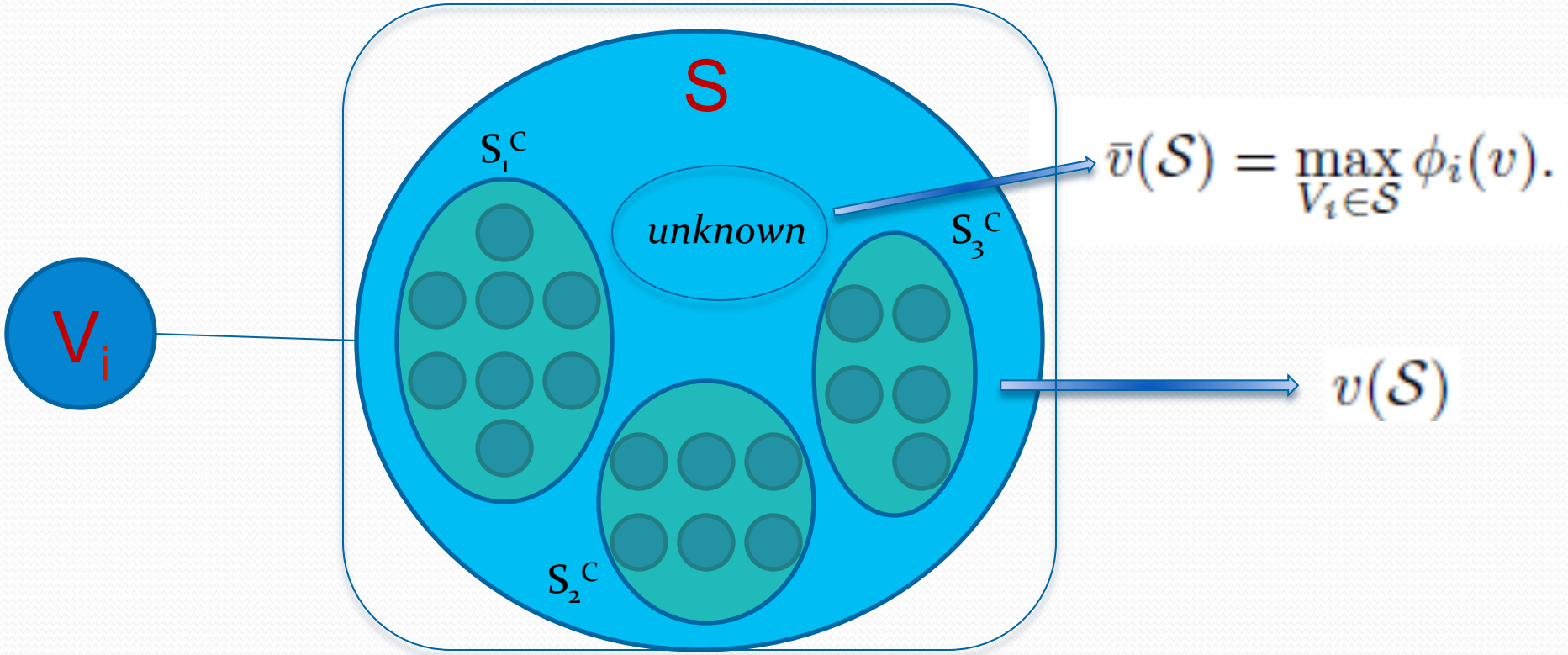
Iterative method

- What if for some set S we have no complete information about the coalitions?



Iterative method

- What if for some set S we have no complete information about the coalitions?



Monotonicity requirement

- Monotonicity of the gain function
  - Bigger coalitions should have higher impacts.
  - Not always the case: e.g., author-publications.
- We impose it using a heuristic.

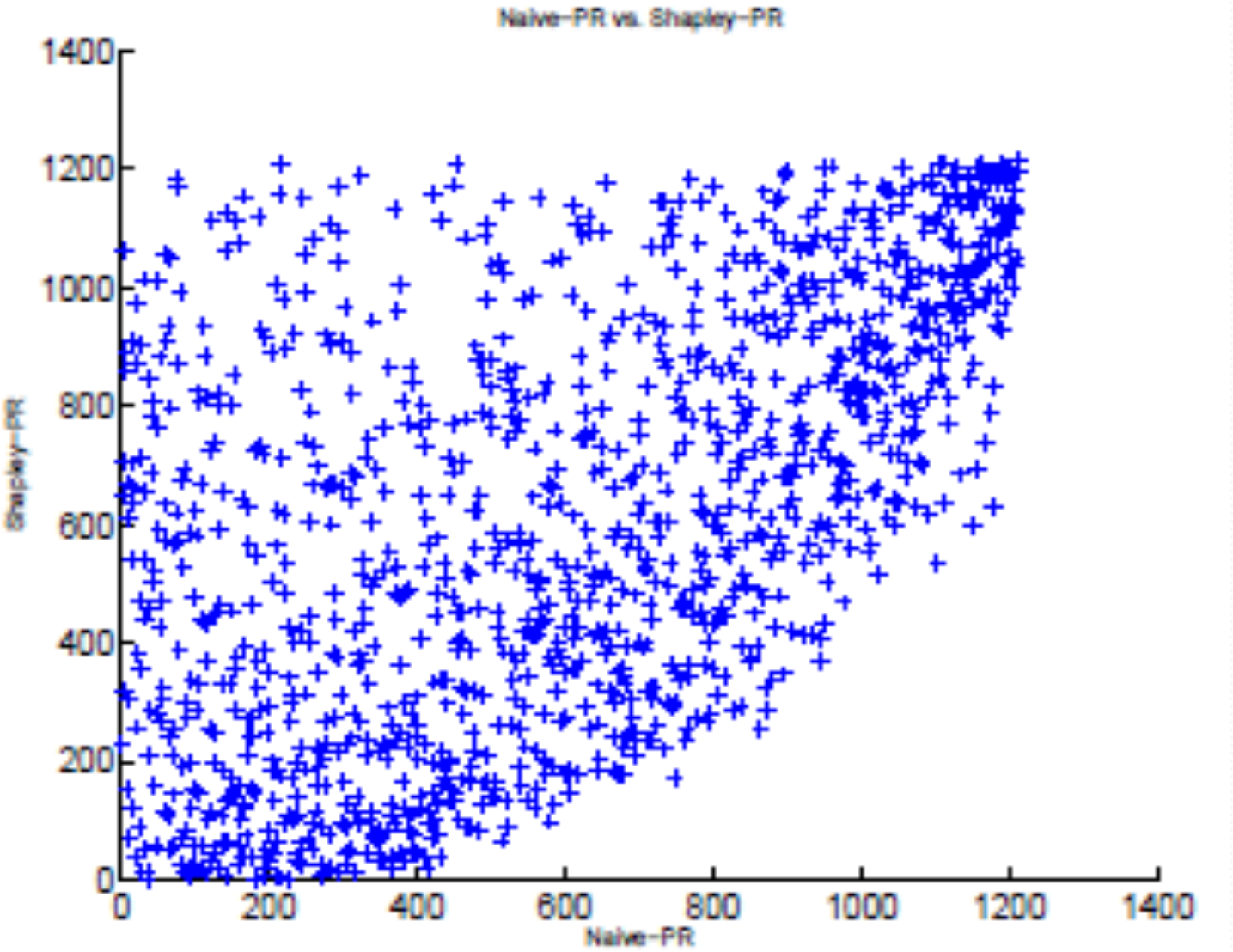
Experiments: setup

- Datasets:
  - ISI Web of Science.
  - Internet Movie Database (IMDB).
- ISI Web of Science:
  - Publication years 2003 and 2009.
  - 1212 authors.

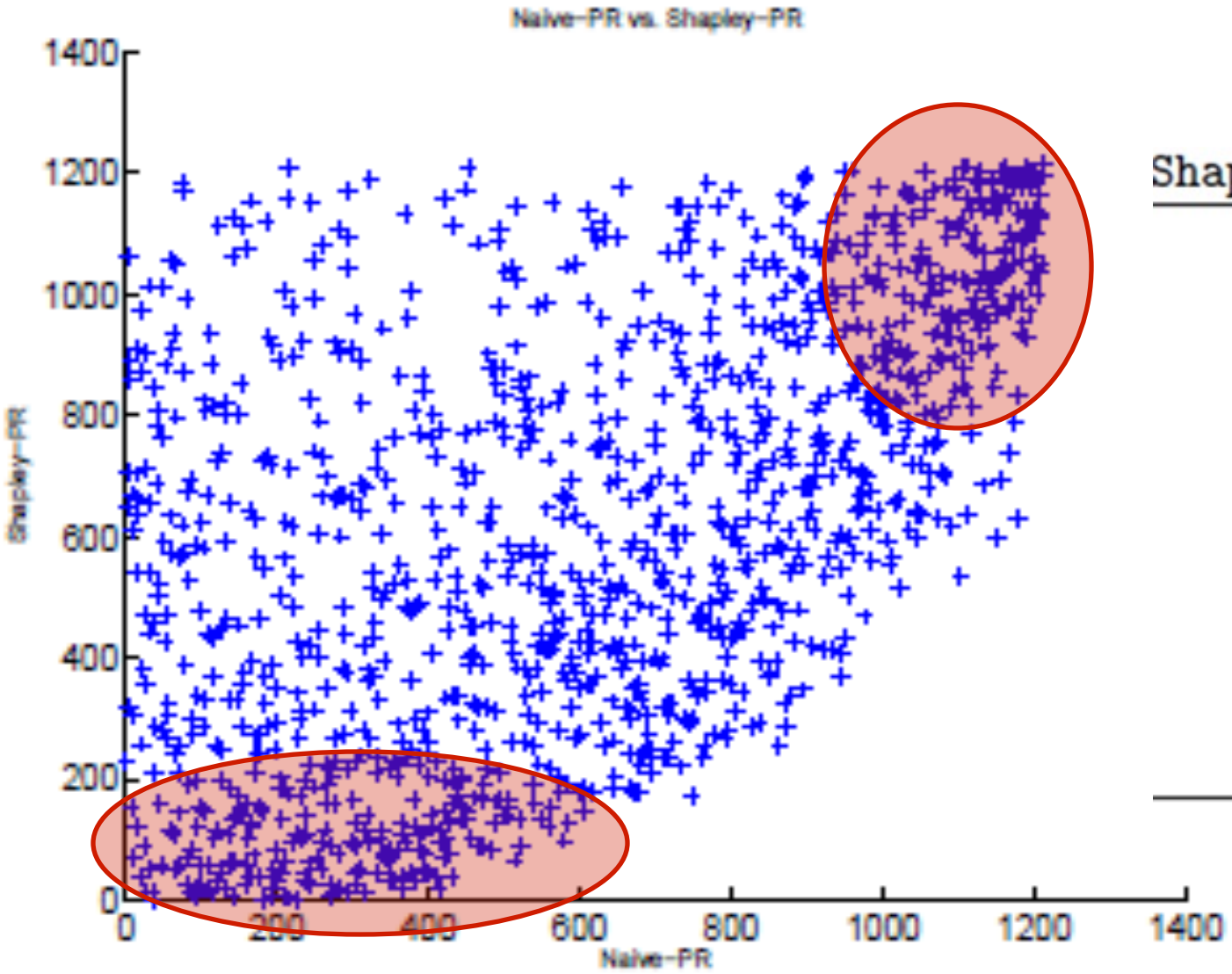
Experiments: setup

- Internet Movie Database:
  - 2000 male actors and 4560 movies.
  - Movie genre type: comedy or action.
  - For each actor we considered only the movies where his credit position was among the top 3.

Naïve PR vs. Shapley PR



Naïve PR vs. Shapley PR



Shapley-PR	Naïve-PR
1	183
2	225
3	35
4	215
5	192
6	272
7	94
8	141
9	208
10	114

Examples

Shapley-PR	Naïve-PR
1	183
2	225
3	35
4	215
5	192
6	272
7	94
8	141
9	208
10	114

Examples

Shapley-PR	Naïve-PR
1	183
2	225
3	35
4	215
5	192
6	272
7	94
8	141
9	208
10	114

P. Papapetrou, Aris Gionis, and Heikki Mannila, "A Shapley value Approach for Influence Attribution" *ECML-PKDD 2011*

Experimental Evaluation

- Top-10 actors given by the Shapley method.

Actor Name	Shapley	Naïve	Actor Name	Naïve	Shapley
Robert De Niro	1	3	Peter Sellers	1	14
Al Pacino	2	8	Jack Nicholson	2	11
Brad Pitt	3	15	Robert De Niro	3	1
Bruce Willis	4	7	Adam Sandler	4	59
Arnold Schwarzenegger	5	24	Daniel Day-Lewis	5	36
Will Smith	6	13	Chris Farley	6	20
Eddie Murphy	7	10	Bruce Willis	7	4
Robin Williams	8	9	Al Pacino	8	2
Morgan Freeman	9	17	Robin Williams	9	8
Ben Stiller	10	29	Eddie Murphy	10	7

Present and Future

- Two main topics:

  Sequence Analysis

  Social networks

Present

- Two main topics:

Sequence Analysis

Social networks

- Influence attribution

ECML-PKDD 2011: P. Papapetrou, A. Gionis, and H. Mannila, "**A Shapley value Approach for Influence Attribution**"

Present

- Two main topics:

  Sequence Analysis — Social networks

  - time series

ACM TODS 2011: P. Papapetrou, V. Athitsos, M. Potamias, G. Kollios, and D. Gunopulos, "**Embedding-based subsequence matching of large time series databases**"

PVLDB 2011: A. Kotsifakos, P. Papapetrou, J. Hollmen, and D. Gunopulos, "**A Ssubsequence Matching with Gaps-Error-Tolerances Framework: a query-by-humming application**"

Present

- Two main topics:

Sequence Analysis

- time series

- event sequences

Social networks

ECML-PKDD 2011: J. Lijffijt, P. Papapetrou, K. Puolamäki, and H. Mannila, "**Analyzing Word Frequencies in Large Text Corpora using Inter-arrival Times and Bootstrapping**"

IJDMB 2011: P. Papapetrou, G.Benson, and G. Kollios, "**Mining Poly-regions in DNA**"

ECML-PKDD 2011: O. Kostakis, P. Papapetrou, and J. Hollmen, "**ARTEMIS: Assessing the Similarity of Event-Interval Sequences**"

Hence, our method is correct!

Shapley-PR	Naïve-PR
1	183
2	225
3	35
4	215
5	192
6	272
7	94
8	141
9	208
10	114

Future

- Two main topics:

  Sequence Analysis

  - burstiness in large texts

  - other domains: DNA?

  - still interested in intervals

  - still interested in music

  Social networks

Future

- Two main topics:

Sequence Analysis

- burstiness in large texts
- other domains: DNA?
- still interested in intervals
- still interested in music

Social networks

- influence attribution
- topic evolution