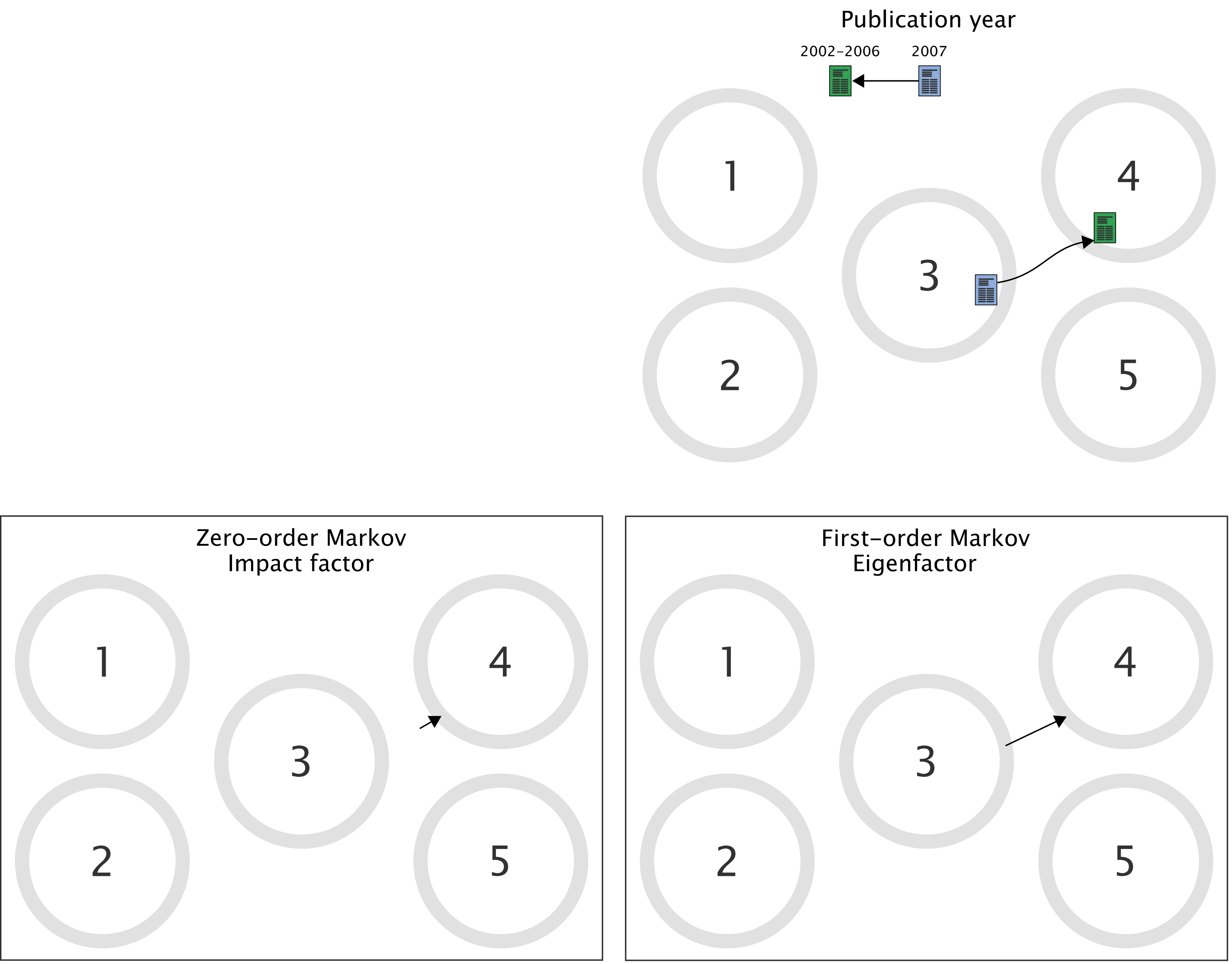
Machine learning for robust rankings

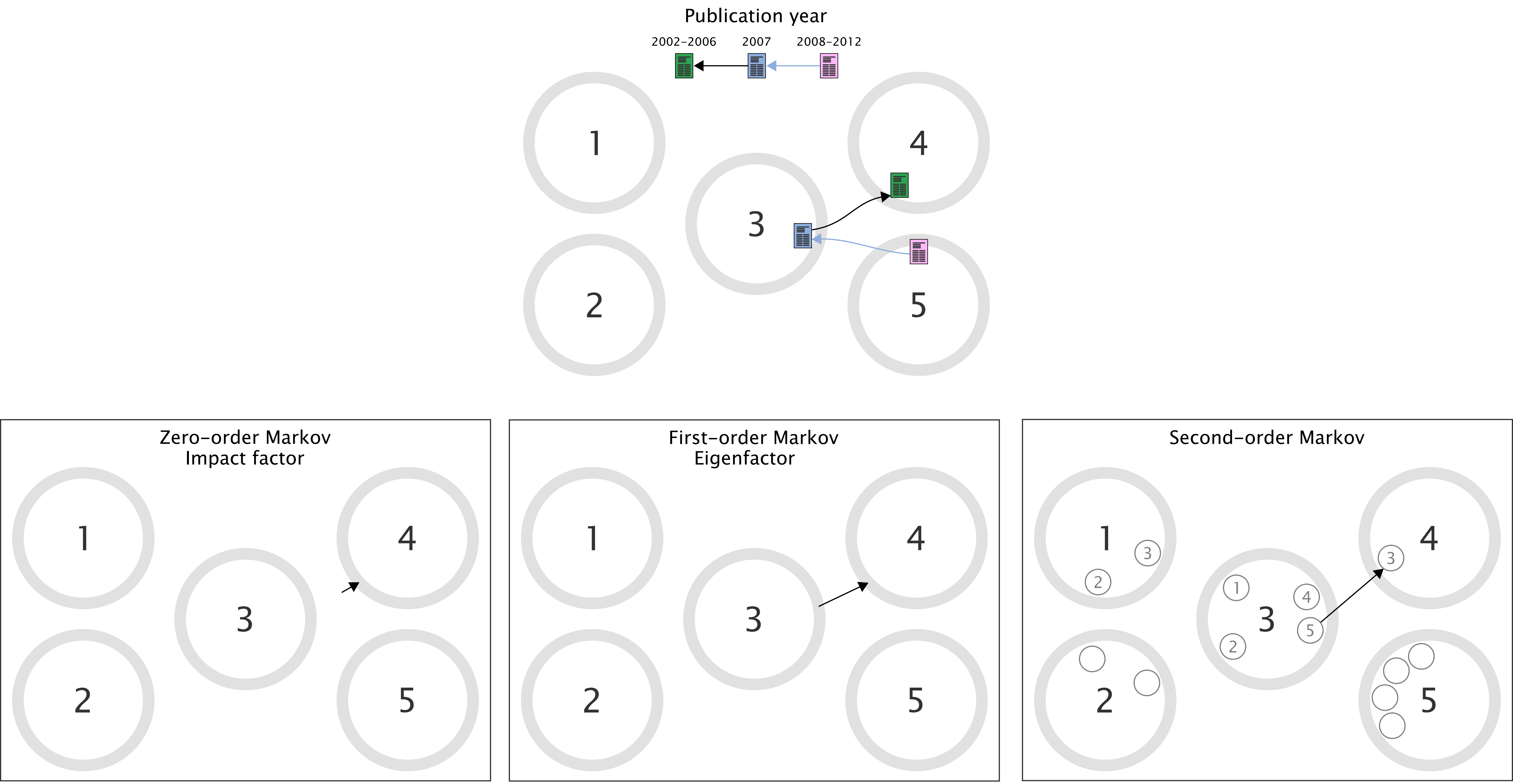
Martin Rosvall

Ludvig Bohlin, Andrea Lancichinetti

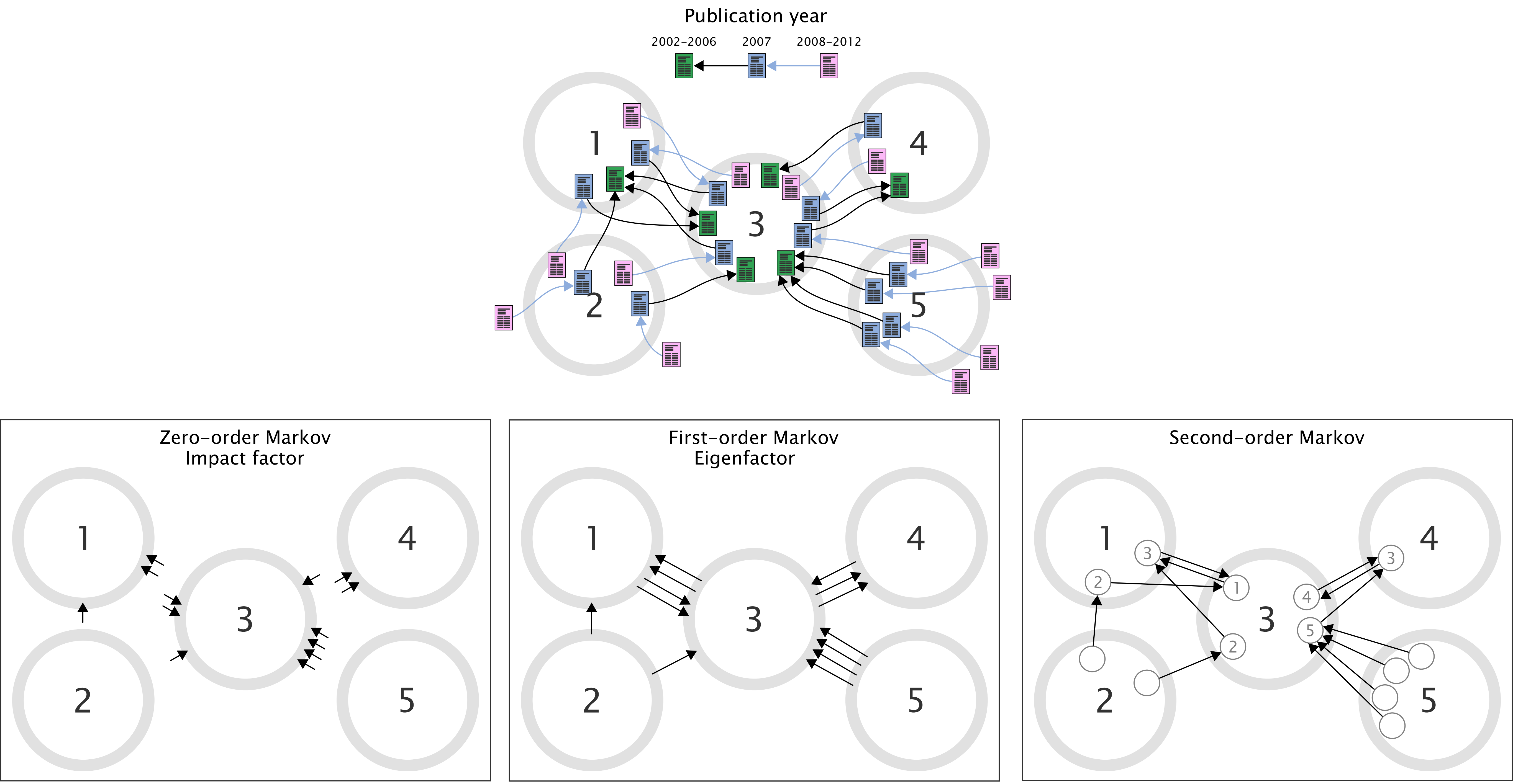
How robust are journal rankings to the selection of journals?

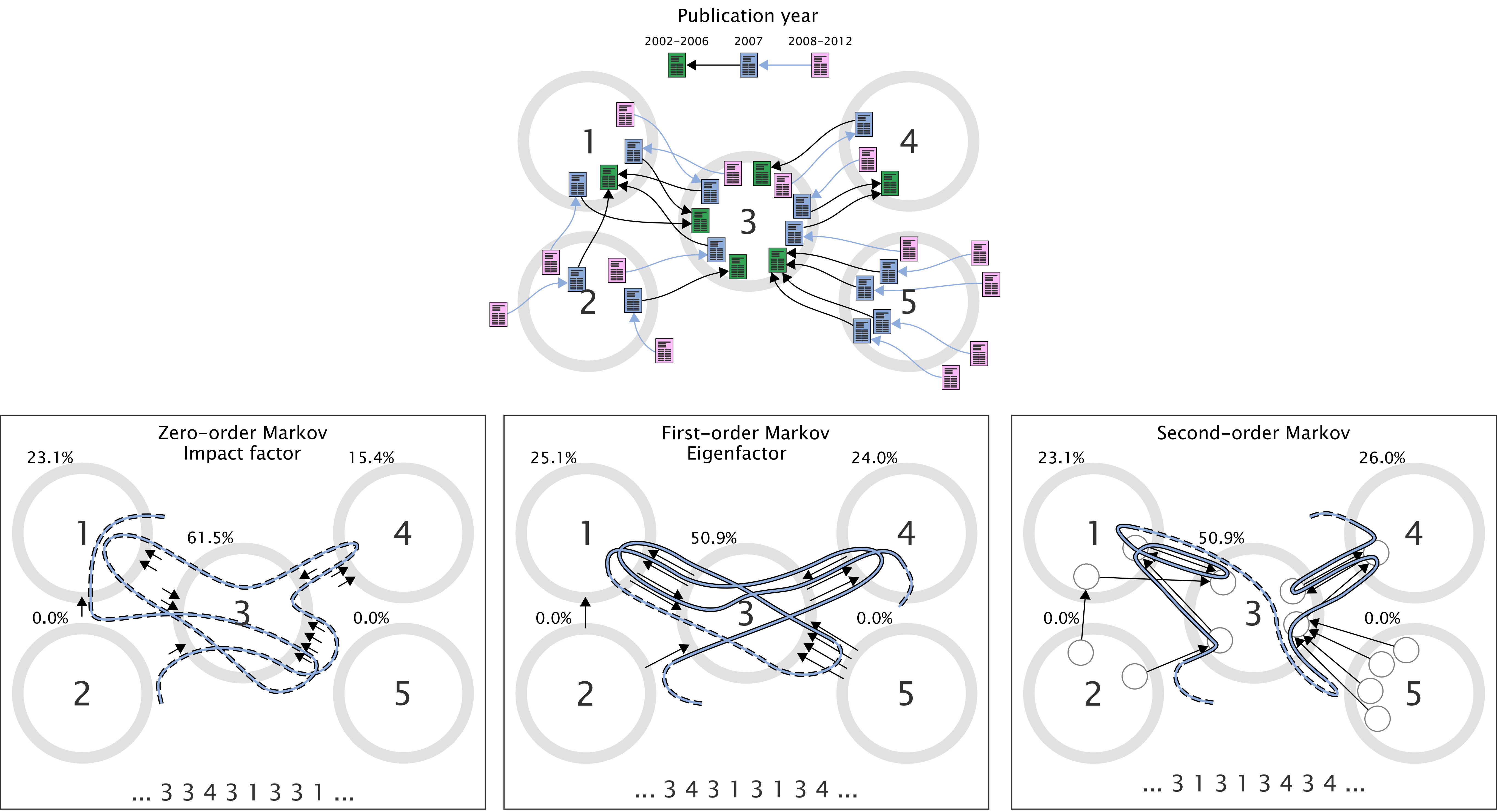


Publication year
2002–2006
2007
1
2
3
4
5
Zero–order Markov
Impact factor
1
2
3
4
5
First–order Markov
Eigenfactor
1
2
3
4
5

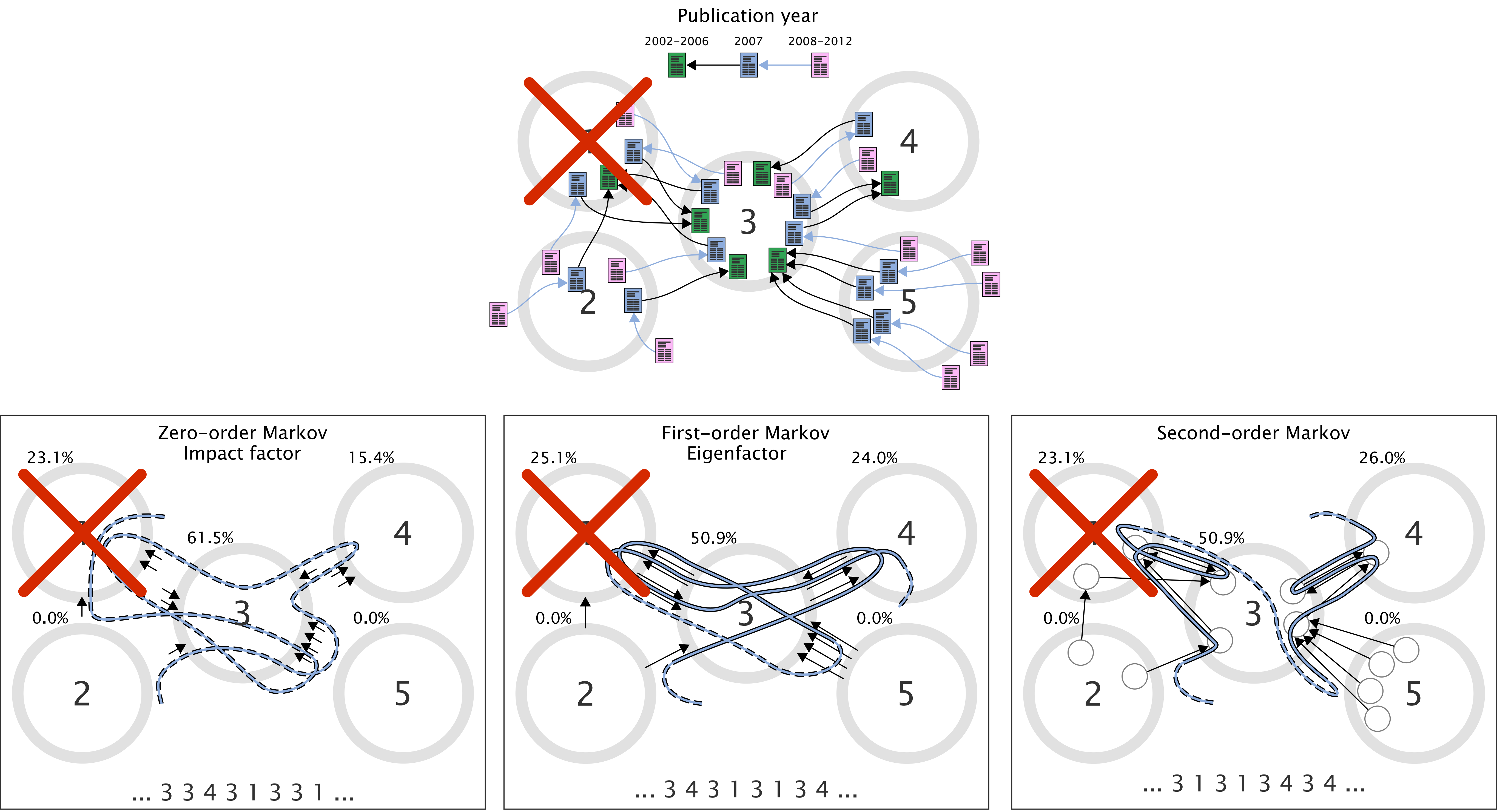


Publication year
2002–2006
2007
2008–2012
1
2
3
4
5
Zero–order Markov
Impact factor
First–order Markov
Eigenfactor
Second–order Markov


Publication year
2002–2006
2007
2008–2012
1
2
3
4
5
Zero–order Markov
Impact factor
1
2
3
4
5
First–order Markov
Eigenfactor
1
2
3
4
5
Second–order Markov
1
2
3
4
5
3
2
1
2
4
5
3

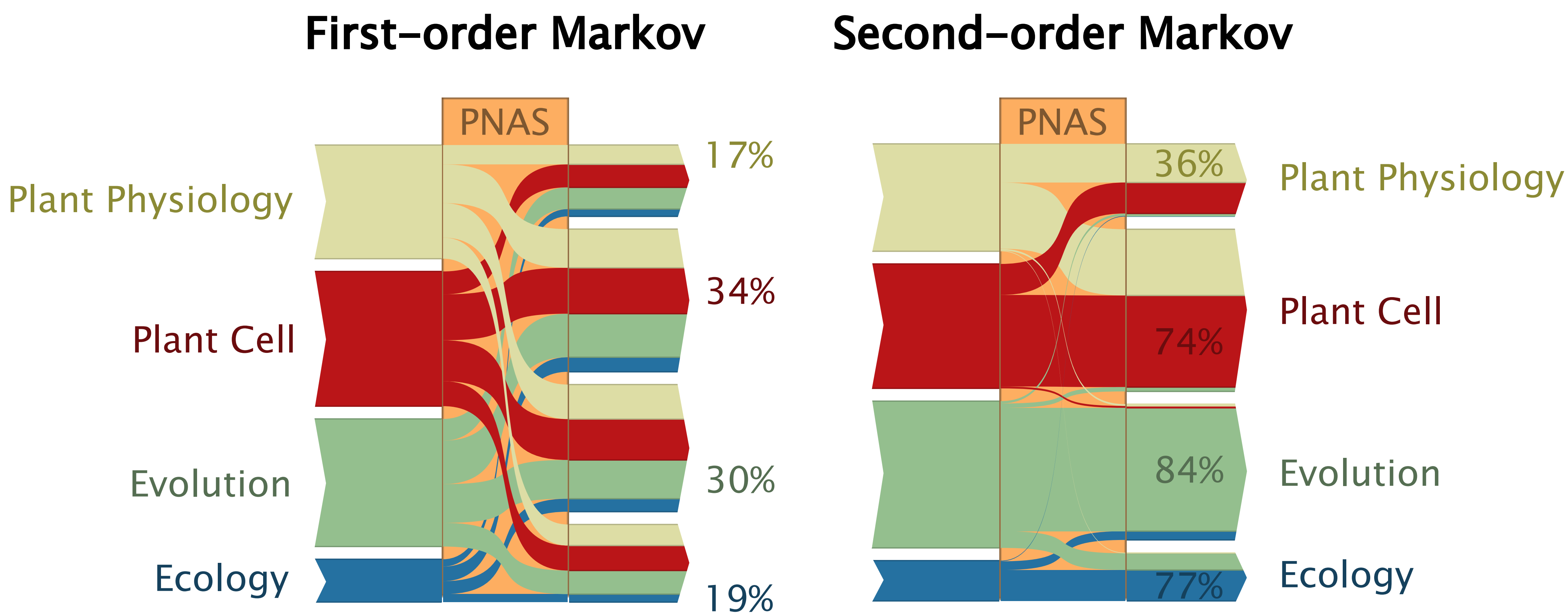


Publication year
2002–2006
2007
2008–2012
1
2
3
4
5
Zero-order Markov
Impact factor
23.1%
15.4%
61.5%
0.0%
0.0%
... 3 3 4 3 1 3 3 1 ...
First-order Markov
Eigenfactor
25.1%
24.0%
50.9%
0.0%
0.0%
... 3 4 3 1 3 1 3 4 ...
Second-order Markov
23.1%
26.0%
50.9%
0.0%
0.0%
... 3 1 3 1 3 4 3 4 ...

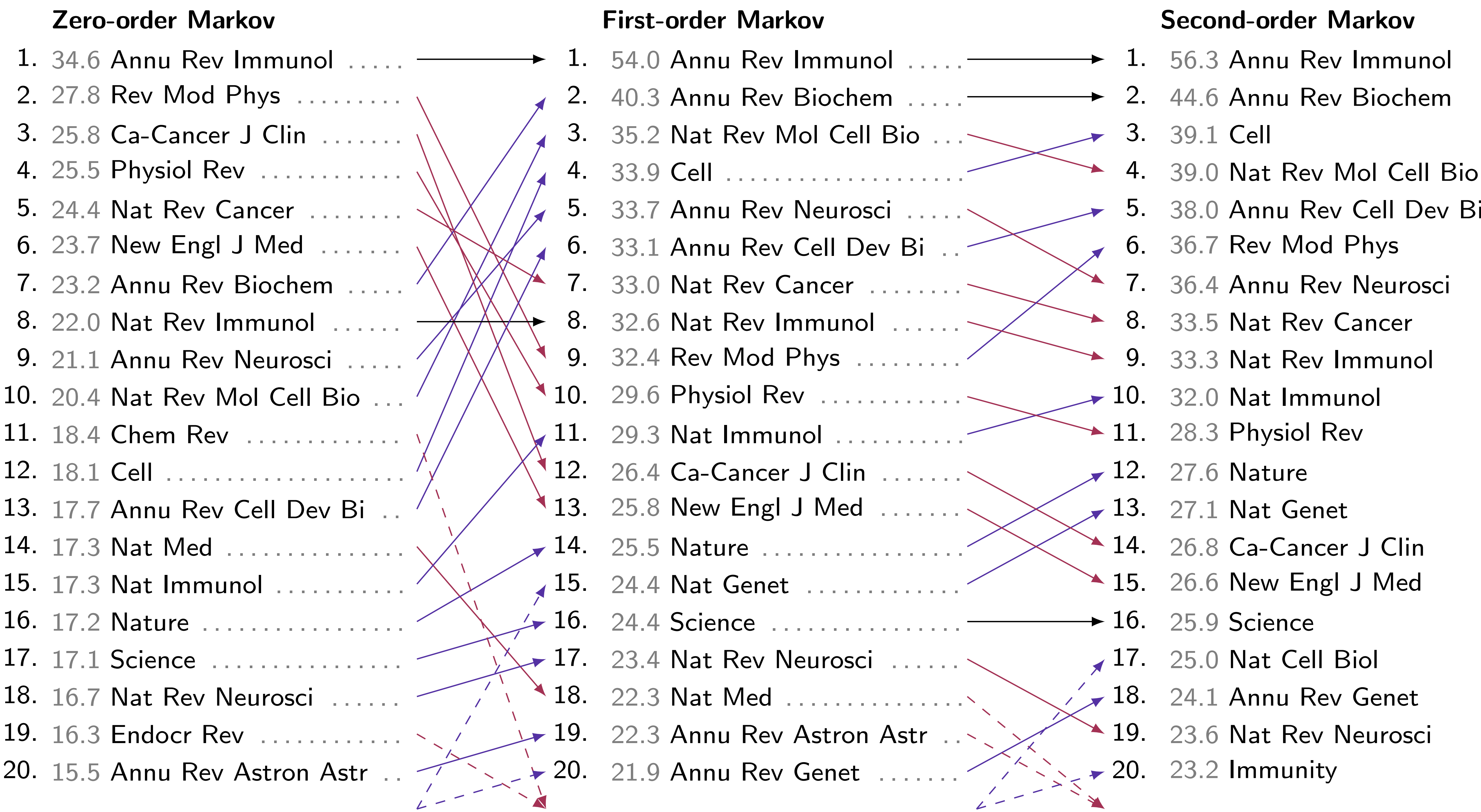


Publication year
2002–2006
2007
2008–2012
3
4
2
5
Zero–order Markov
Impact factor
23.1%
15.4%
61.5%
0.0%
0.0%
... 3 3 4 3 1 3 3 1 ...
First–order Markov
Eigenfactor
25.1%
24.0%
50.9%
0.0%
0.0%
... 3 4 3 1 3 1 3 4 ...
Second–order Markov
23.1%
26.0%
50.9%
0.0%
0.0%
... 3 1 3 1 3 4 3 4 ...

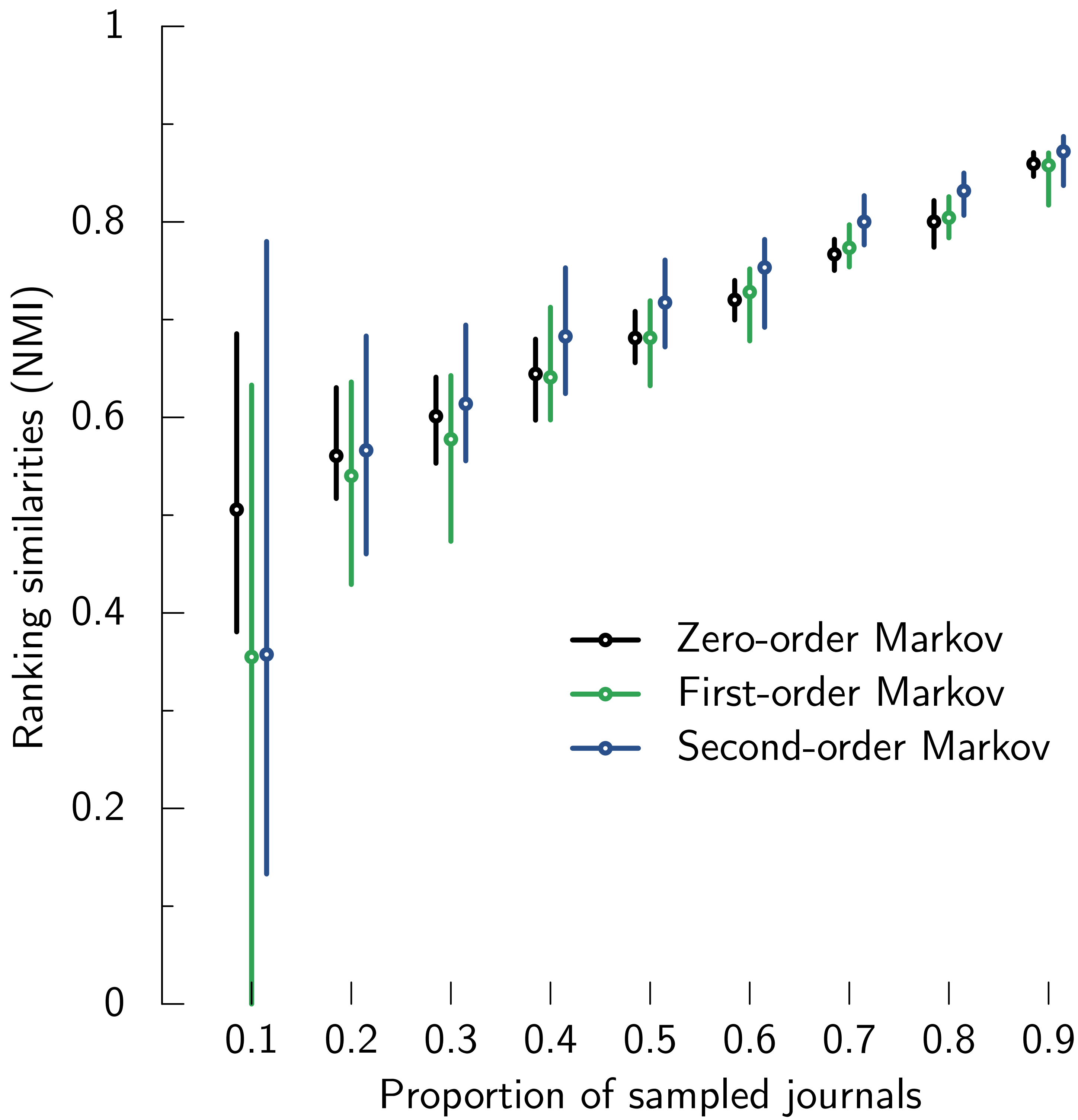
How does the Markov order affect the robustness?



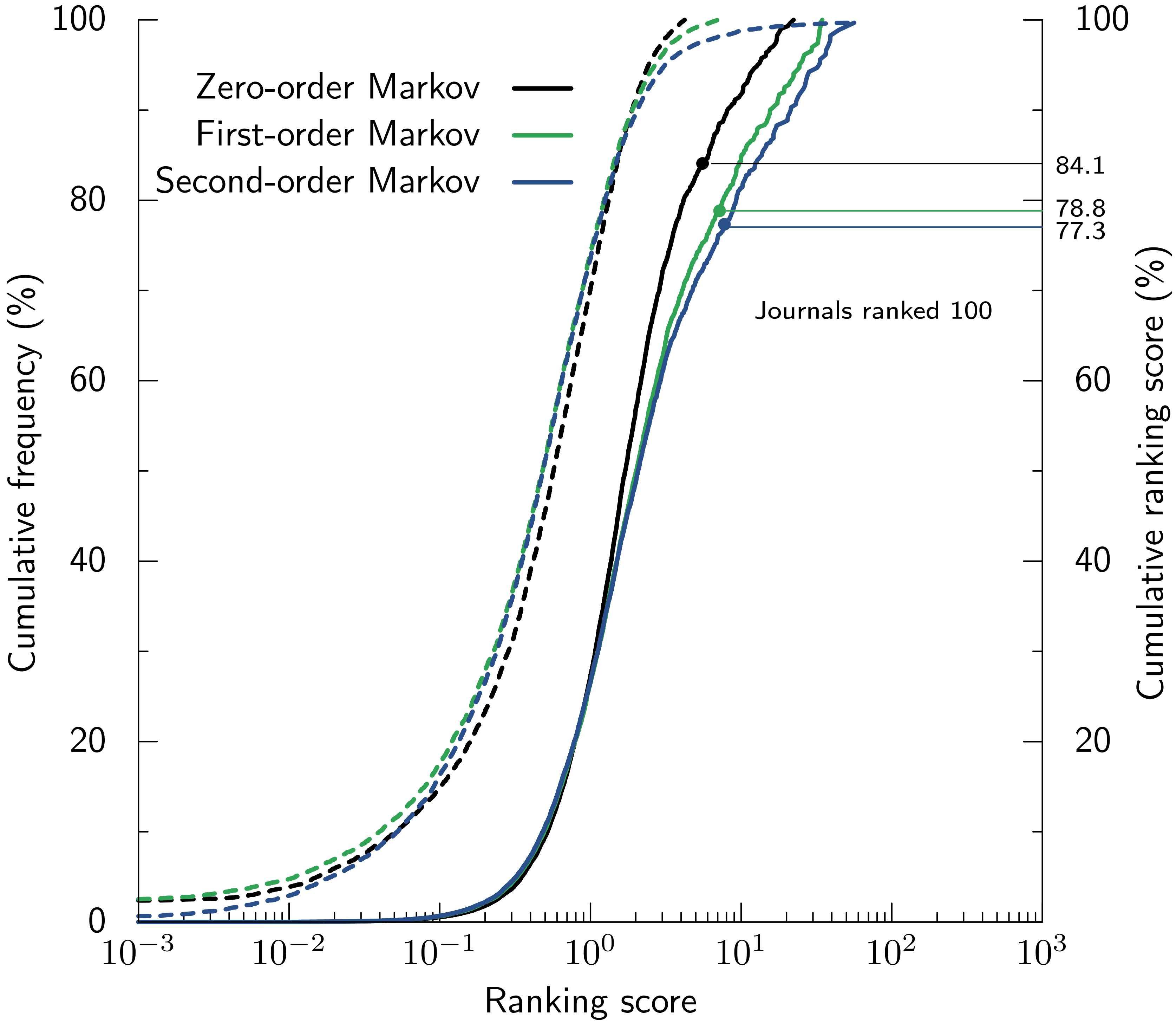
First–order Markov
Second–order Markov
PNAS
Plant Physiology
Plant Cell
Evolution
Ecology
17%
34%
30%
19%
PNAS
36%
74%
84%
77%
Plant Physiology
Plant Cell
Evolution
Ecology



Zero-order Markov
1. 34.6 Annu Rev Immunol
2. 27.8 Rev Mod Phys
3. 25.8 Ca-Cancer J Clin
4. 25.5 Physiol Rev
5. 24.4 Nat Rev Cancer
6. 23.7 New Engl J Med
7. 23.2 Annu Rev Biochem
8. 22.0 Nat Rev Immunol
9. 21.1 Annu Rev Neurosci
10. 20.4 Nat Rev Mol Cell Bio
11. 18.4 Chem Rev
12. 18.1 Cell
13. 17.7 Annu Rev Cell Dev Bi
14. 17.3 Nat Med
15. 17.3 Nat Immunol
16. 17.2 Nature
17. 17.1 Science
18. 16.7 Nat Rev Neurosci
19. 16.3 Endocr Rev
20. 15.5 Annu Rev Astron Astr
First-order Markov
1. 54.0 Annu Rev Immunol
2. 40.3 Annu Rev Biochem
3. 35.2 Nat Rev Mol Cell Bio
4. 33.9 Cell
5. 33.7 Annu Rev Neurosci
6. 33.1 Annu Rev Cell Dev Bi
7. 33.0 Nat Rev Cancer
8. 32.6 Nat Rev Immunol
9. 32.4 Rev Mod Phys
10. 29.6 Physiol Rev
11. 29.3 Nat Immunol
12. 26.4 Ca-Cancer J Clin
13. 25.8 New Engl J Med
14. 25.5 Nature
15. 24.4 Nat Genet
16. 24.4 Science
17. 23.4 Nat Rev Neurosci
18. 22.3 Nat Med
19. 22.3 Annu Rev Astron Astr
20. 21.9 Annu Rev Genet
Second-order Markov
1. 56.3 Annu Rev Immunol
2. 44.6 Annu Rev Biochem
3. 39.1 Cell
4. 39.0 Nat Rev Mol Cell Bio
5. 38.0 Annu Rev Cell Dev Bi
6. 36.7 Rev Mod Phys
7. 36.4 Annu Rev Neurosci
8. 33.5 Nat Rev Cancer
9. 33.3 Nat Rev Immunol
10. 32.0 Nat Immunol
11. 28.3 Physiol Rev
12. 27.6 Nature
13. 27.1 Nat Genet
14. 26.8 Ca-Cancer J Clin
15. 26.6 New Engl J Med
16. 25.9 Science
17. 25.0 Nat Cell Biol
18. 24.1 Annu Rev Genet
19. 23.6 Nat Rev Neurosci
20. 23.2 Immunity



Ranking similarities (NMI)
0
0.2
0.4
0.6
0.8
1
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Proportion of sampled journals
Zero-order Markov
First-order Markov
Second-order Markov


Zero-order Markov
First-order Markov
Second-order Markov
84.1
78.8
77.3
Journals ranked 100
Cumulative frequency (%)
Cumulative ranking score (%)
Ranking score
0
20
40
60
80
100
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}

Conclusion

Higher orders give more robust rankings because

+ the range increases
+ citation weights depend on journal importance
+ perturbations remain local
+ higher predictability in cross-validation test

– at the cost of requiring more data

Remark

Second-order Markov dynamics are essential for identifying multidisciplinary journals

Field	PNAS		Science		Ecology		Plant Cell	
	M1	M2	M1	M2	M1	M2	M1	M2
Ecology	-	13	-	29	100	100	-	-
Cell biology	100	80	100	68	-	-	100	100
Mathematics	-	4.6	-	-	-	-	-	-
Statistics	-	1.5	-	-	-	-	-	-
Anthropology	-	-	-	1.6	-	-	-	-
Others	-	0.38 (1)	-	1.4 (7)	-	-	-	-

Robustness of journal rankings by network flows with different amounts of memory

Ludvig Bohlin, Alcides Viamontes Esquivel, Andrea Lancichinetti, Martin Rosvall

(Submitted on 30 May 2014)

As the number of scientific journals has multiplied, journal rankings have become increasingly important for scientific decisions. From submissions and subscriptions to grants and hirings, researchers, policy makers, and funding agencies make important decisions with influence from journal rankings such as the ISI journal impact factor. Typically, the rankings are derived from the citation network between a selection of journals and unavoidably depend on this selection. However, little is known about how robust rankings are to the selection of included journals. Here we compare the robustness of three journal rankings based on network flows induced on citation networks. They model pathways of researchers navigating scholarly literature, stepping between journals and remembering their previous steps to different degree: zero-step memory as impact factor, one-step memory as Eigenfactor, and two-step memory, corresponding to zero-, first-, and second-order Markov models of citation flow between journals. We conclude that a second-order Markov model is slightly more robust, because it combines the advantages of the lower-order models: perturbations that remain local and citation weights that depend on journal importance. However, the robustness gain comes at the cost of requiring more data, because the second-order Markov model requires citation data from twice as long a period.

Comments:	8 pages, 5 figures
Subjects:	**Physics and Society (physics.soc-ph)**; Digital Libraries (cs.DL)
Cite as:	arXiv:1405.7832 **[physics.soc-ph]**
	(or **arXiv:1405.7832v1** **[physics.soc-ph]** for this version)

Submission history

From: Ludvig Bohlin [view email]
[v1] Fri, 30 May 2014 12:09:58 GMT (408kb,D)

nature COMMUNICATIONS

ARTICLE

Received 18 Mar 2014 | Accepted 9 Jul 2014 | Published 11 Aug 2014

DOI: 10.1038/ncomms5630

Memory in network flows and its effects on spreading dynamics and community detection

Martin Rosvall[1], Alcides V. Esquivel[1], Andrea Lancichinetti[1,2], Jevin D. West[1,3] & Renaud Lambiotte[4]

Random walks on networks is the standard tool for modelling spreading processes in social and biological systems. This first-order Markov approach is used in conventional community detection, ranking and spreading analysis, although it ignores a potentially important feature of the dynamics: where flow moves to may depend on where it comes from. Here we analyse pathways from different systems, and although we only observe marginal consequences for disease spreading, we show that ignoring the effects of second-order Markov dynamics has important consequences for community detection, ranking and information spreading. For example, capturing dynamics with a second-order Markov model allows us to reveal actual travel patterns in air traffic and to uncover multidisciplinary journals in scientific communication. These findings were achieved only by using more available data and making no additional assumptions, and therefore suggest that accounting for higher-order memory in network flows can help us better understand how real systems are organized and function.