

# On Differences in the Tagging Behavior of Spammers and Regular Users – Supplemental Material

Klaas Dellschaft  
 Department of Computer Science  
 Universität Koblenz-Landau  
 Koblenz, Germany  
 klaasd@uni-koblenz.de

Steffen Staab  
 Department of Computer Science  
 Universität Koblenz-Landau  
 Koblenz, Germany  
 staab@uni-koblenz.de

## 1. INTRODUCTION

This supplemental material gives the detailed plots for all co-occurrence stream pairs which have been used in [1]. Each stream pair consists of a *filtered stream* and an *unfiltered stream*. The filtered stream only contains tag assignments made by regular users while the unfiltered streams contain tag assignments made by regular users as well as spammers. More details about the research background and the data set are available in [1].

The statistics of the stream pairs are given in Tab. 2 and 1. For each of the stream pairs, we give the following information:

- The parameters for which the Epistemic Model and the Yule-Simon Model with memory generated achieved the best fit to the tag frequency distribution in the filtered stream.
- A plot which shows the empirical distribution function of the filtered stream and compares it to the best fitting empirical distribution function generated with the Epistemic Model and the Yule-Simon Model with Memory.
- The empirical distribution function of the tag frequency distribution in the filtered and unfiltered stream.
- The frequency-rank plot of the tag frequency distribution in the filtered and unfiltered stream.
- A plot of the vocabulary growth in the filtered and unfiltered stream.
- The frequency-rank plot of the word frequency distribution in a web corpus which has a similar topical focus as the stream pair (see [1] for more details).

The raw data sets and the software for simulating the Epistemic Model as well as the Yule-Simon Model with Memory are available from <http://west.uni-koblenz.de/Research/DataSets/websci2010>.

## 2. REFERENCES

- [1] K. Dellschaft and S. Staab. On differences in the tagging behavior of spammers and regular users. In *Proceedings of the Web Science Conference 2010*, 2010. <http://www.isweb.uni-koblenz.de/files/publications/Dellschaft2010ODI.pdf>.

Copyright is held by the authors.

*Web Science Conf. 2010*, April 26-27, 2010, Raleigh, NC, USA.

tag	Users	Tag Assignments
ringtones	3,215	74,155
setup	4,176	40,689
boat	1,641	23,512
historical	1,374	16,662
messages	973	9,634
decorative	223	8,892
costs	709	7,359
ff	482	5,114
checkbox	869	4,758
datawarehouse	444	3,730
tools	183	25,437
social	246	15,322
design	209	14,606
analysis	142	12,506
blogs	85	8,926

**Table 1: Statistics of the filtered streams from Delicious (top) and Bibsonomy (bottom). They contain only tag assignments made by regular users.**

tag	Users		Tag Assignments	
	regular	spammer	regular	spammer
ringtones	1,512	57	23,217	50,938
setup	680	18	6,389	34,300
boat	816	64	16,180	7,332
historical	285	13	3,158	13,504
messages	131	12	965	8,669
decorative	147	21	8,195	697
costs	81	3	441	6,918
ff	469	12	4,935	179
checkbox	112	2	717	4,041
datawarehouse	424	2	3,583	147
tools	114	169	17,729	7,708
social	131	162	9,095	6,227
design	84	44	6,414	8,192
analysis	72	53	9,562	2,944
blogs	50	96	4,768	4,158

**Table 2: Statistics of the unfiltered streams from Delicious (top) and Bibsonomy (bottom). They contain tag assignments of regular users as well as spammers.**

### 3. RINGTONES STREAM PAIR

For the filtered ringtones stream, the best fit of the Epistemic Model has been achieved with the following parameters:

- $I = 0.95$ ;  $n = 1750$ ;  $h = 13000$

The best fit of the Yule-Simon Model with Memory has been achieved with the following parameters:

- $p = 0.072$ ;  $\tau = 480$ ;  $n_0 = 100$

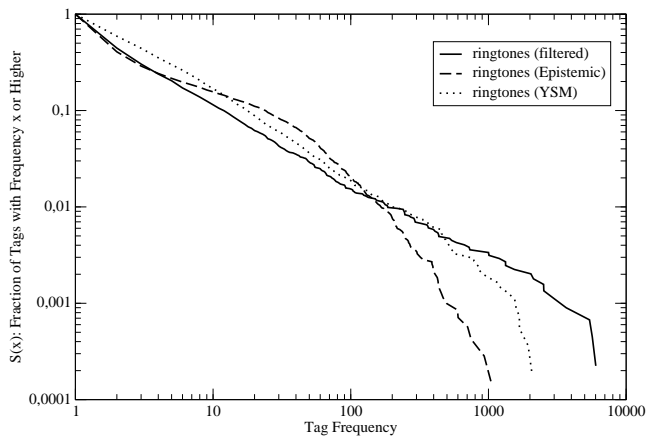


Figure 1: Empirical distribution functions for the filtered *ringtones* and the best fitting graph simulated with the Epistemic Model and the Yule-Simon Model with Memory (YSM).

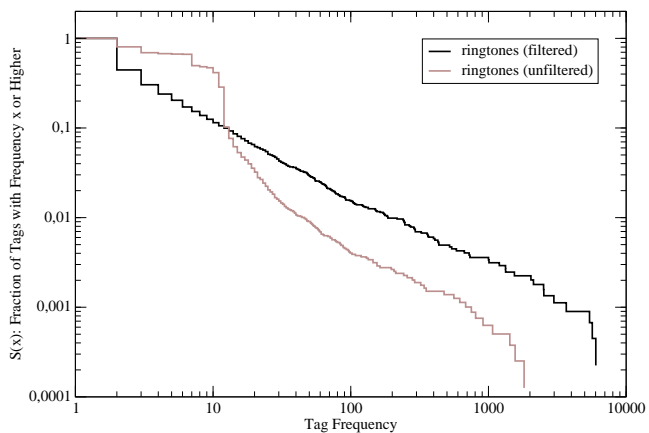


Figure 2: Empirical distribution functions for the *ringtones* stream pair.

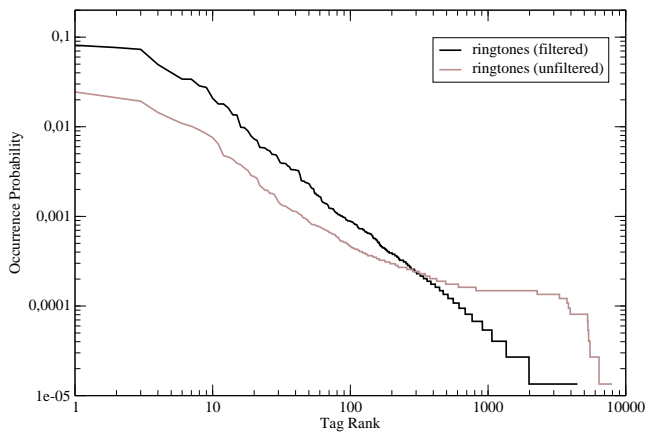


Figure 3: Frequency-rank plot for the *ringtones* stream pair.

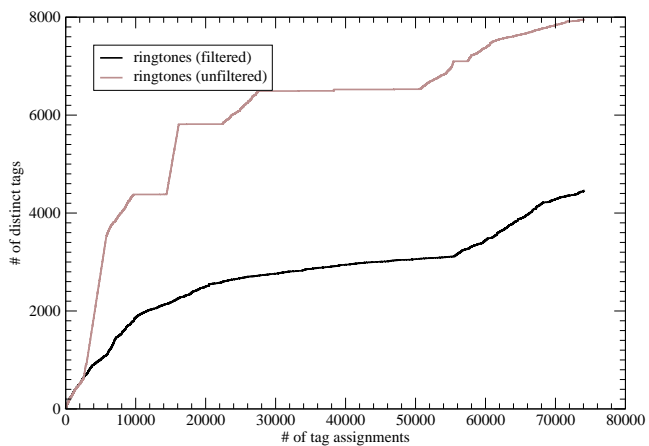


Figure 4: Vocabulary growth for the *ringtones* stream pair.

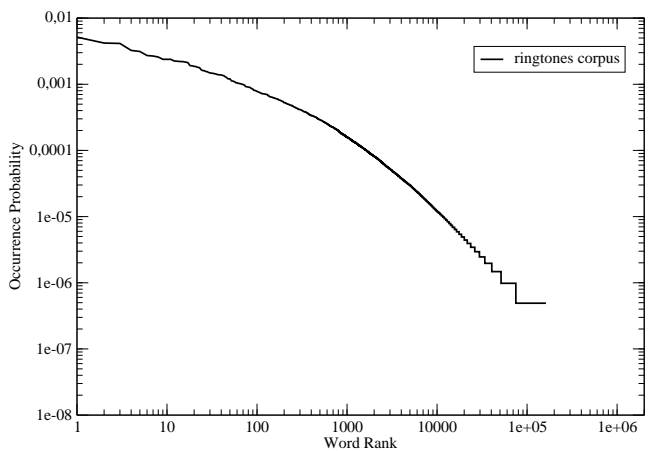


Figure 5: Frequency-rank plot for the *ringtones* web corpus.

## 4. SETUP STREAM PAIR

For the filtered setup stream, the best fit of the Epistemic Model has been achieved with the following parameters:

- $I = 0.682$ ;  $n = 2750$ ;  $h = 11000$

The best fit of the Yule-Simon Model with Memory has been achieved with the following parameters:

- $p = 0.13$ ;  $\tau = 500$ ;  $n_0 = 100$

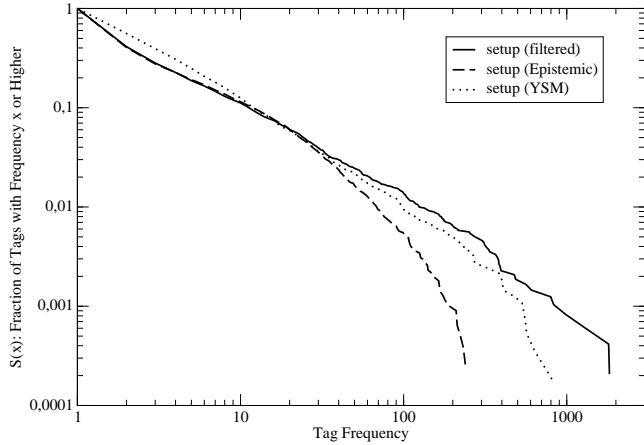


Figure 6: Empirical distribution functions for the filtered *setup* and the best fitting graph simulated with the Epistemic Model and the Yule-Simon Model with Memory (YSM).

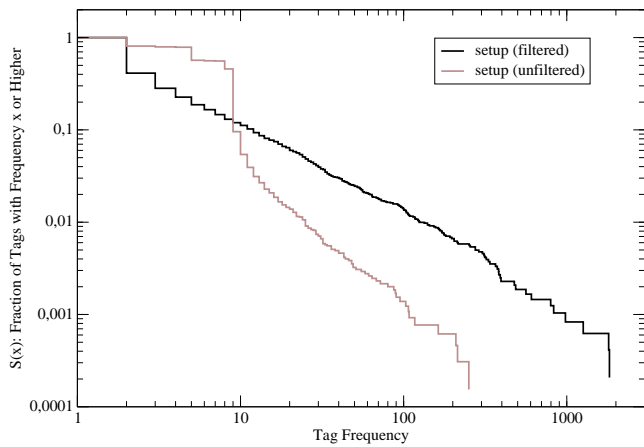


Figure 7: Empirical distribution functions for the *setup* stream pair.

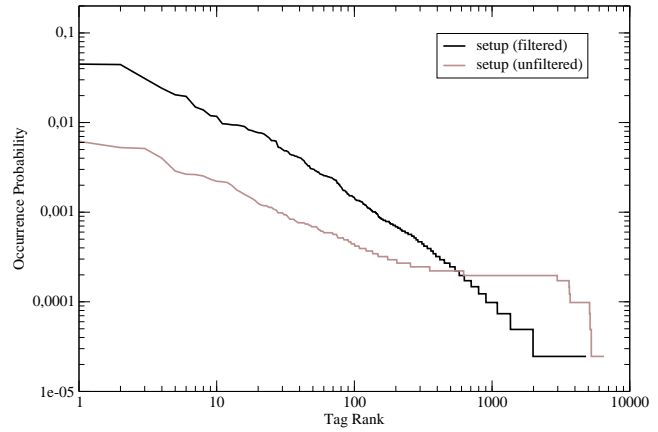


Figure 8: Frequency-rank plot for the *setup* stream pair.

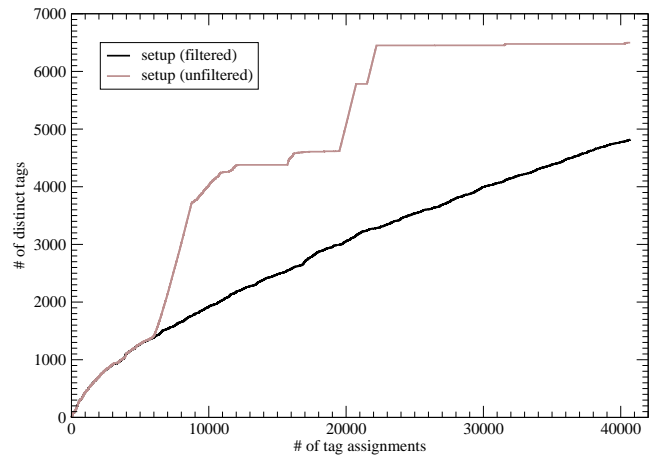


Figure 9: Vocabulary growth for the *setup* stream pair.

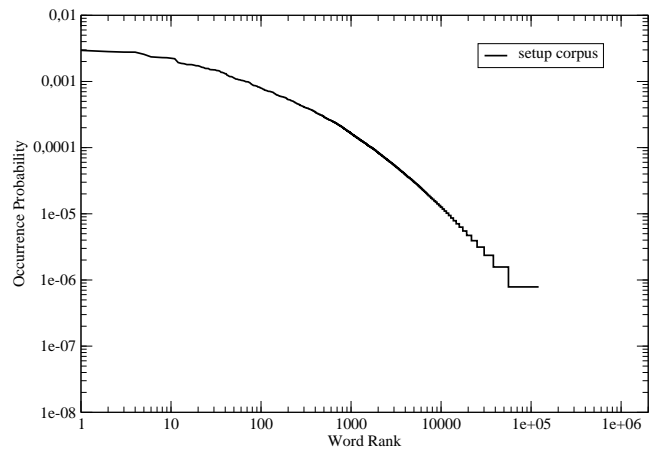


Figure 10: Frequency-rank plot for the *setup* web corpus.

## 5. BOAT STREAM PAIR

For the filtered boat stream, the best fit of the Epistemic Model has been achieved with the following parameters:

- $I = 0.553$ ;  $n = 1850$ ;  $h = 5000$

The best fit of the Yule-Simon Model with Memory has been achieved with the following parameters:

- $p = 0.245$ ;  $\tau = 480$ ;  $n_0 = 100$

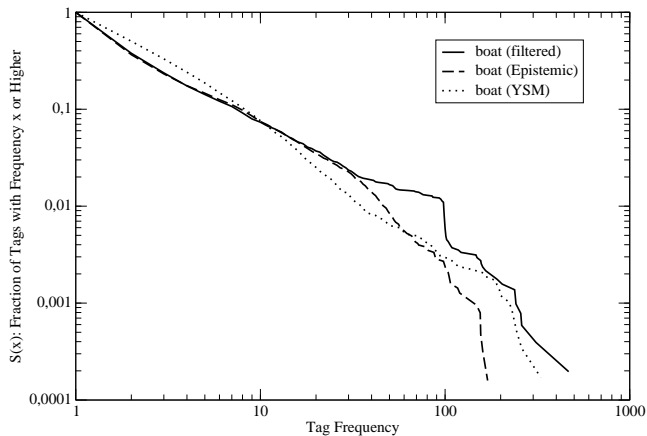


Figure 11: Empirical distribution functions for the filtered *boat* and the best fitting graph simulated with the Epistemic Model and the Yule-Simon Model with Memory (YSM).

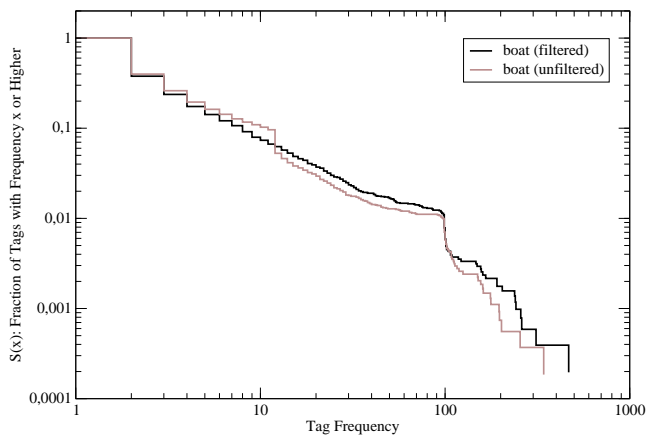


Figure 12: Empirical distribution functions for the *boat* stream pair.

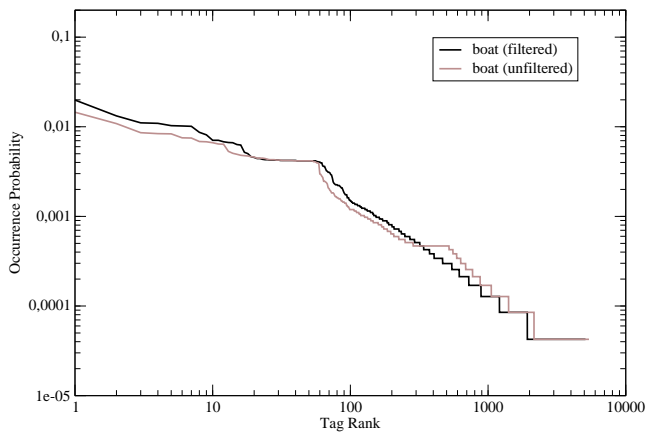


Figure 13: Frequency-rank plot for the *boat* stream pair.

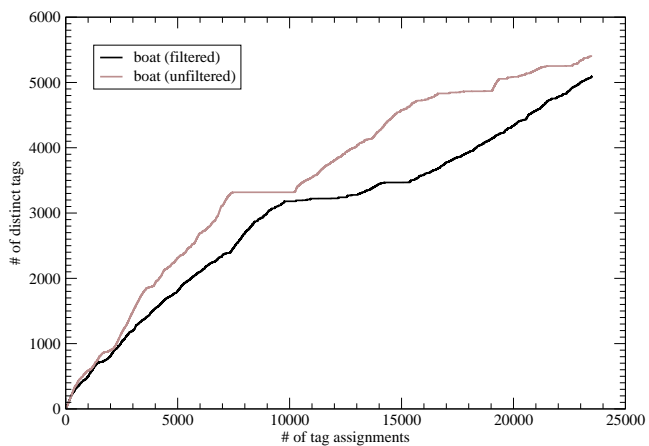


Figure 14: Vocabulary growth for the *boat* stream pair.

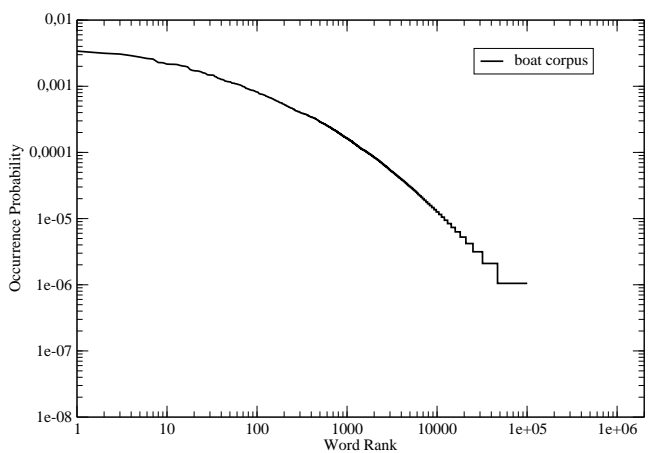


Figure 15: Frequency-rank plot for the *boat* web corpus.

## 6. HISTORICAL STREAM PAIR

For the filtered historical stream, the best fit of the Epistemic Model has been achieved with the following parameters:

- $I = 0.653$ ;  $n = 1000$ ;  $h = 3000$

The best fit of the Yule-Simon Model with Memory has been achieved with the following parameters:

- $p = 0.308$ ;  $\tau = 500$ ;  $n_0 = 100$

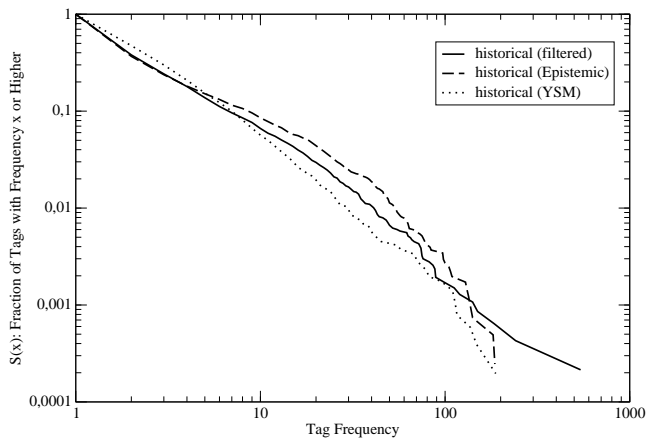


Figure 16: Empirical distribution functions for the filtered *historical* and the best fitting graph simulated with the Epistemic Model and the Yule-Simon Model with Memory (YSM).

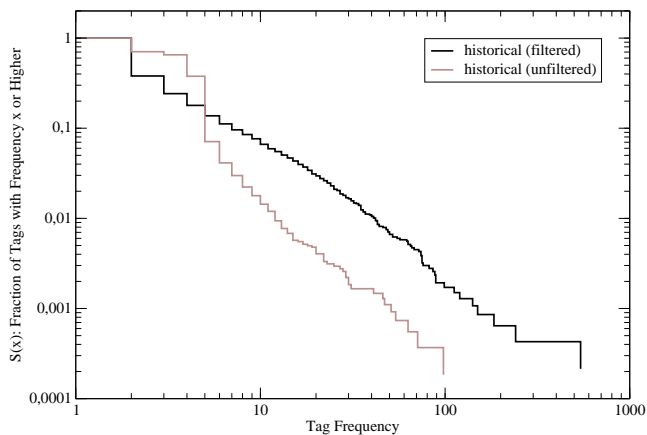


Figure 17: Empirical distribution functions for the *historical* stream pair.

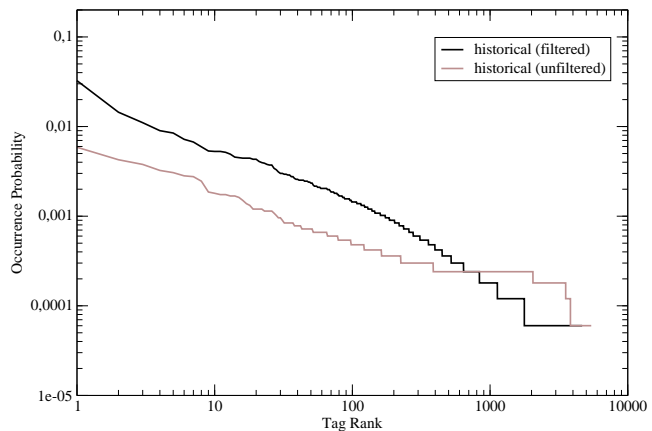


Figure 18: Frequency-rank plot for the *historical* stream pair.

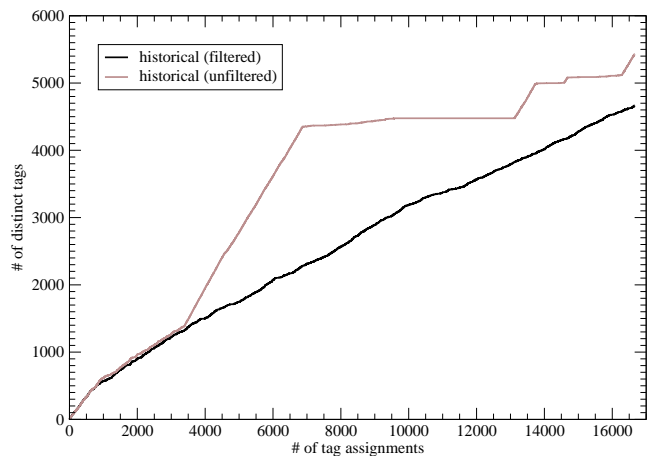


Figure 19: Vocabulary growth for the *historical* stream pair.

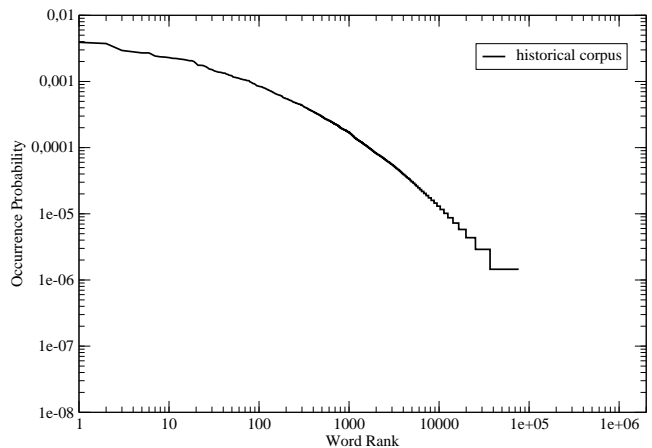


Figure 20: Frequency-rank plot for the *historical* web corpus.

## 7. MESSAGES STREAM PAIR

For the filtered messages stream, the best fit of the Epistemic Model has been achieved with the following parameters:

- $I = 0.501$ ;  $n = 500$ ;  $h = 1000$

The best fit of the Yule-Simon Model with Memory has been achieved with the following parameters:

- $p = 0.353$ ;  $\tau = 500$ ;  $n_0 = 100$

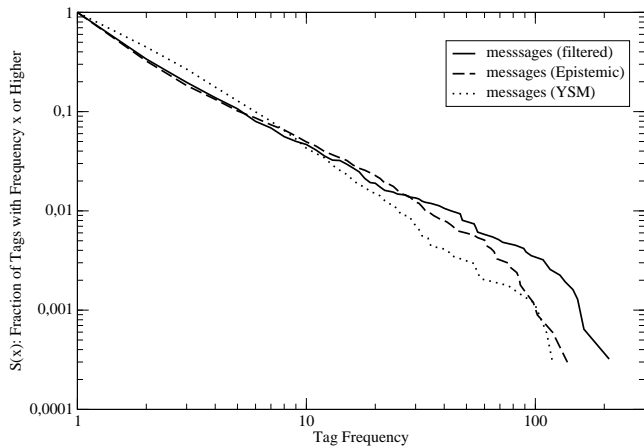


Figure 21: Empirical distribution functions for the filtered *messages* and the best fitting graph simulated with the Epistemic Model and the Yule-Simon Model with Memory (YSM).

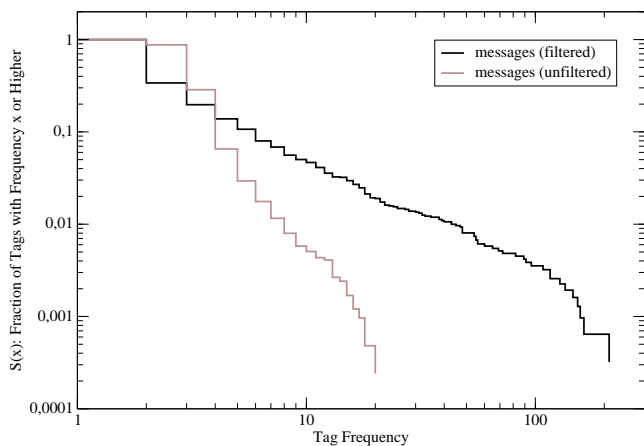


Figure 22: Empirical distribution functions for the *messages* stream pair.

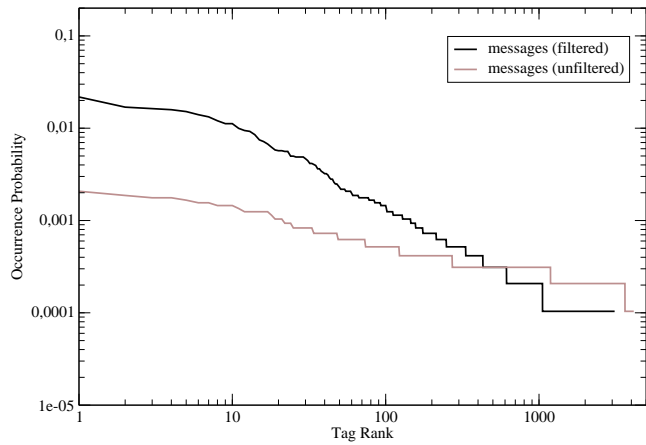


Figure 23: Frequency-rank plot for the *messages* stream pair.

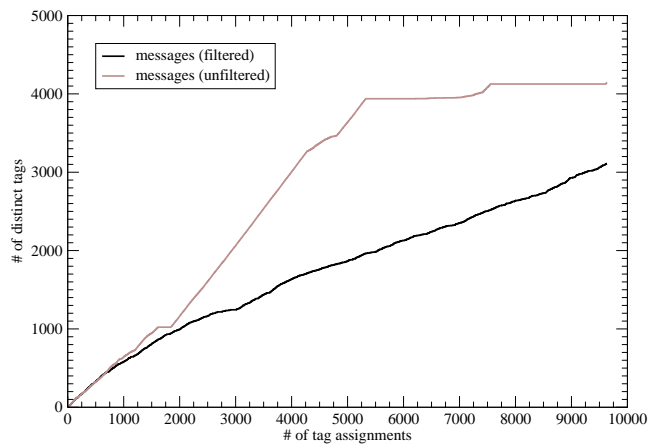


Figure 24: Vocabulary growth for the *messages* stream pair.

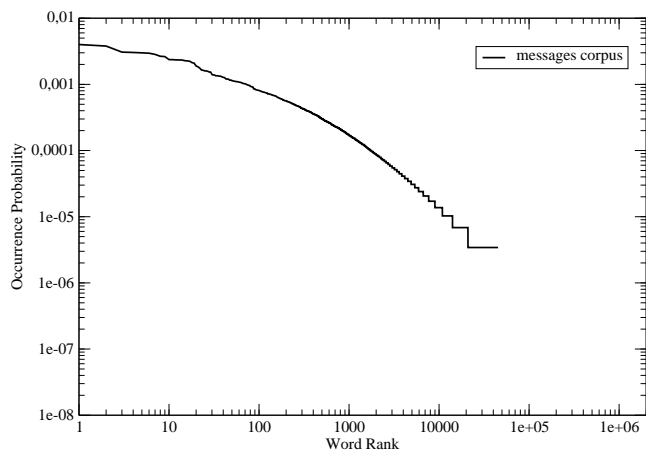


Figure 25: Frequency-rank plot for the *messages* web corpus.

## 8. DECORATIVE STREAM PAIR

For the filtered decorative stream, the best fit of the Epistemic Model has been achieved with the following parameters:

- $I = 0.95$ ;  $n = 300$ ;  $h = 3000$

The best fit of the Yule-Simon Model with Memory has been achieved with the following parameters:

- $p = 0.14$ ;  $\tau = 500$ ;  $n_0 = 100$

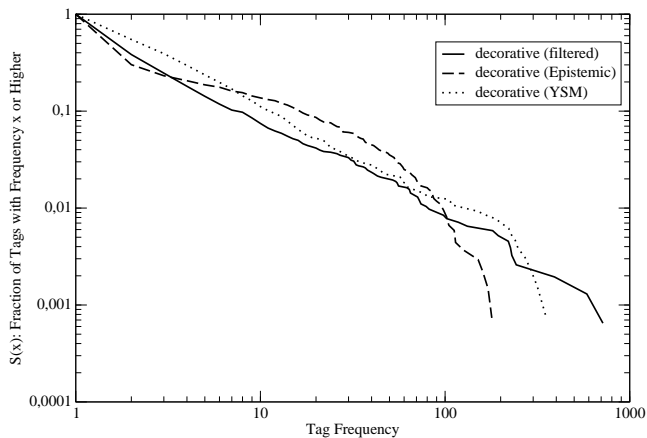


Figure 26: Empirical distribution functions for the filtered *decorative* and the best fitting graph simulated with the Epistemic Model and the Yule-Simon Model with Memory (YSM).

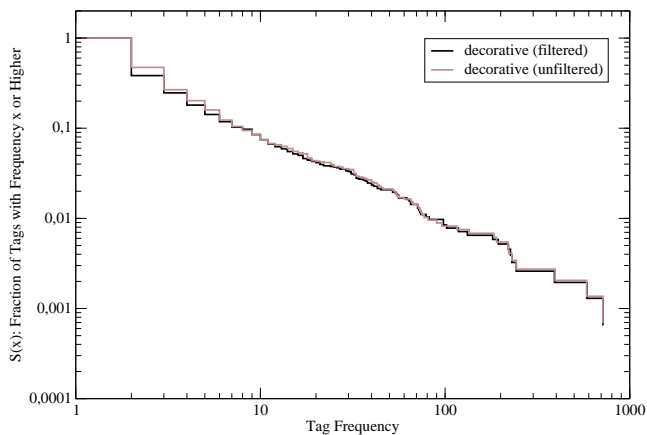


Figure 27: Empirical distribution functions for the *decorative* stream pair.

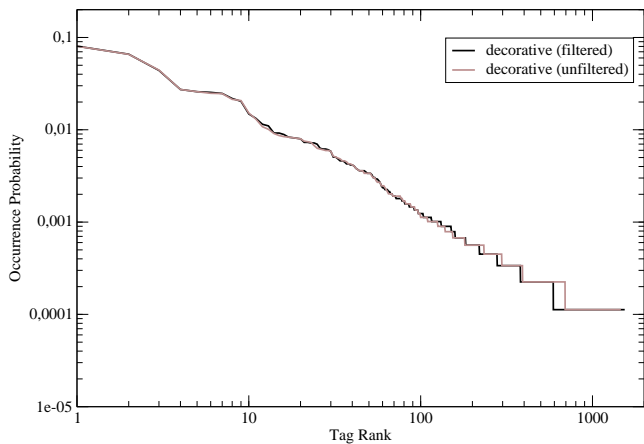


Figure 28: Frequency-rank plot for the *decorative* stream pair.

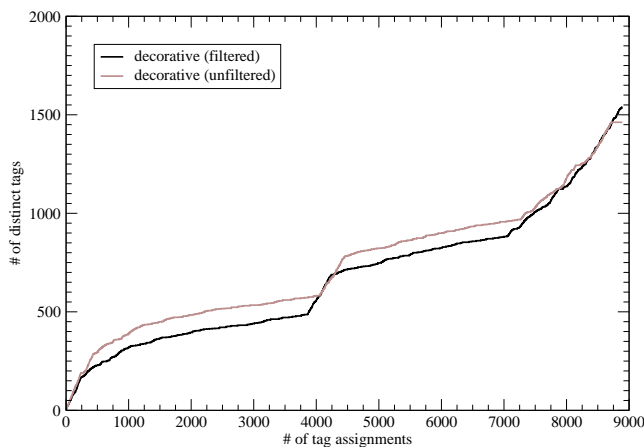


Figure 29: Vocabulary growth for the *decorative* stream pair.

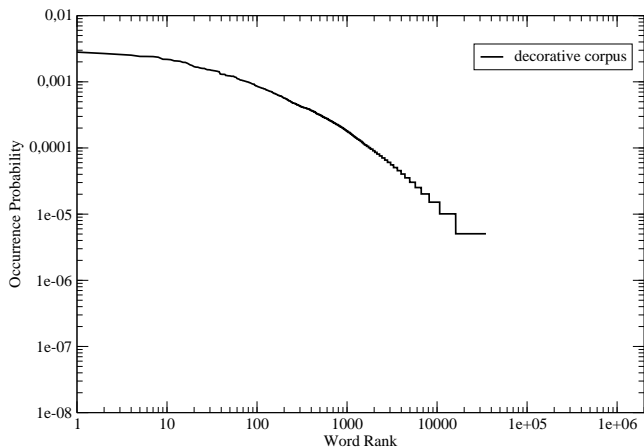


Figure 30: Frequency-rank plot for the *decorative* web corpus.

## 9. COSTS STREAM PAIR

For the filtered costs stream, the best fit of the Epistemic Model has been achieved with the following parameters:

- $I = 0.888$ ;  $n = 200$ ;  $h = 1000$

The best fit of the Yule-Simon Model with Memory has been achieved with the following parameters:

- $p = 0.38$ ;  $\tau = 500$ ;  $n_0 = 100$

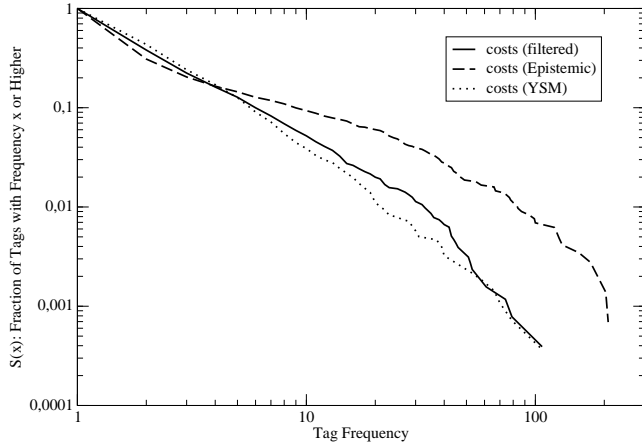


Figure 31: Empirical distribution functions for the filtered *costs* and the best fitting graph simulated with the Epistemic Model and the Yule-Simon Model with Memory (YSM).

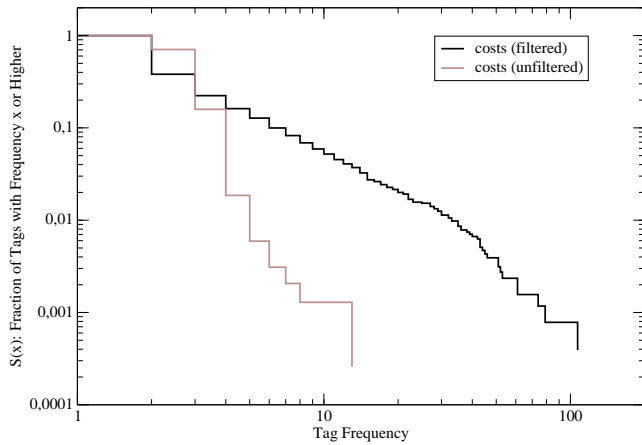


Figure 32: Empirical distribution functions for the *costs* stream pair.

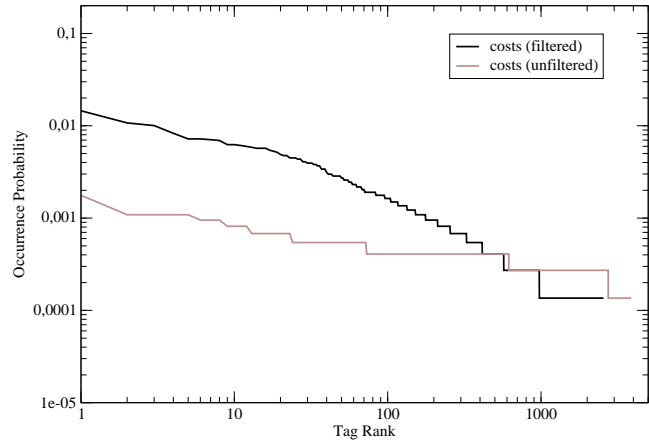


Figure 33: Frequency-rank plot for the *costs* stream pair.

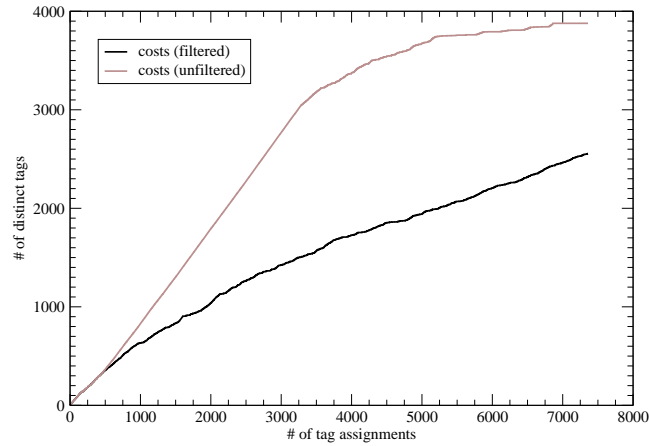


Figure 34: Vocabulary growth for the *costs* stream pair.

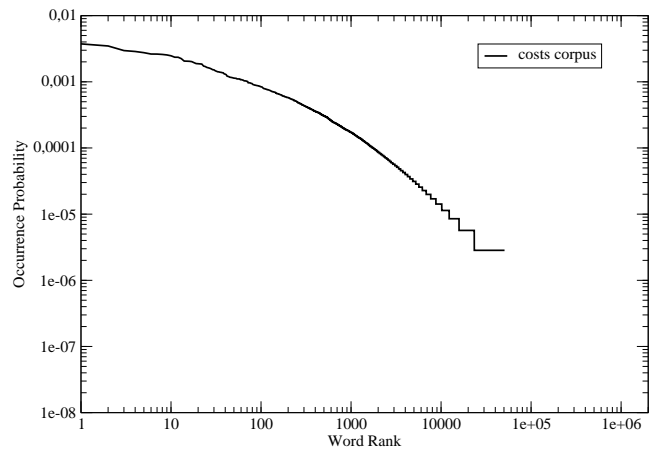


Figure 35: Frequency-rank plot for the *costs* web corpus.

## 10. FF STREAM PAIR

For the filtered  $ff$  stream, the best fit of the Epistemic Model has been achieved with the following parameters:

- $I = 0.597$ ;  $n = 1350$ ;  $h = 5000$

The best fit of the Yule-Simon Model with Memory has been achieved with the following parameters:

- $p = 0.322$ ;  $\tau = 480$ ;  $n_0 = 100$

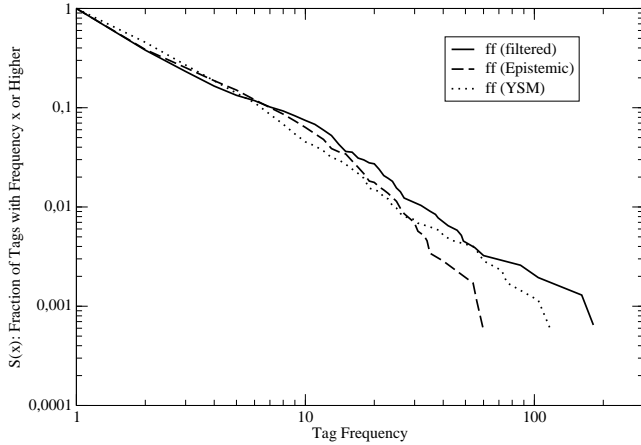


Figure 36: Empirical distribution functions for the filtered  $ff$  and the best fitting graph simulated with the Epistemic Model and the Yule-Simon Model with Memory (YSM).

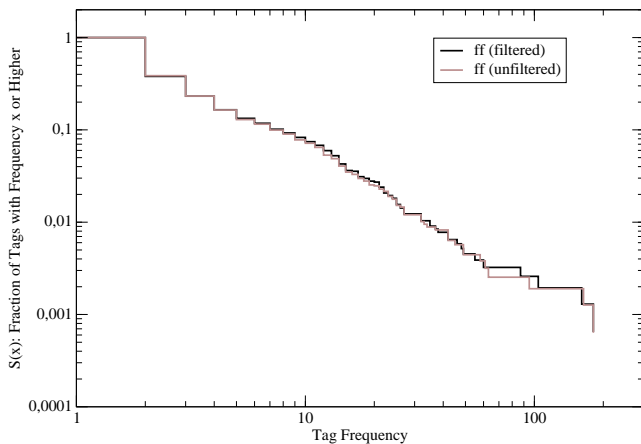


Figure 37: Empirical distribution functions for the  $ff$  stream pair.

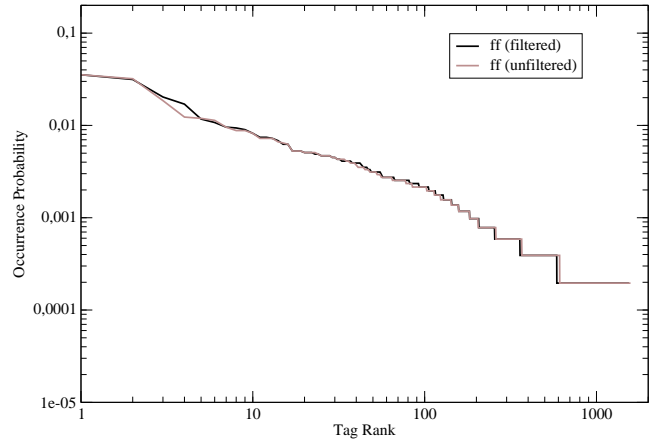


Figure 38: Frequency-rank plot for the  $ff$  stream pair.

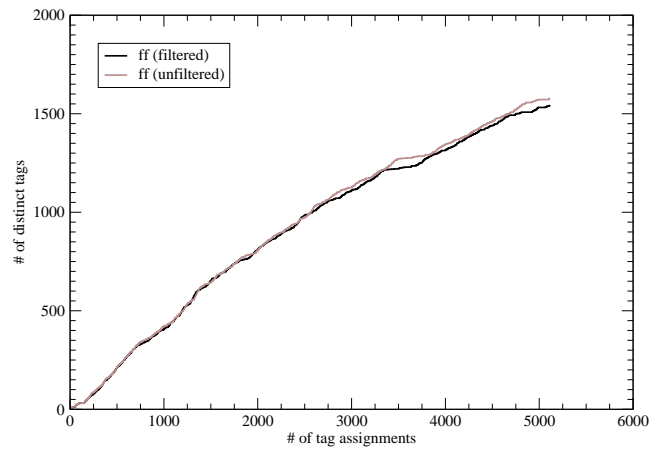


Figure 39: Vocabulary growth for the  $ff$  stream pair.

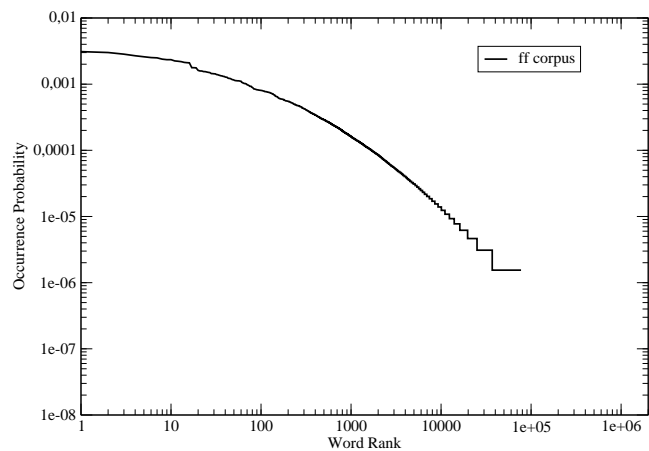


Figure 40: Frequency-rank plot for the  $ff$  web corpus.

## 11. CHECKBOX STREAM PAIR

For the filtered checkbox stream, the best fit of the Epistemic Model has been achieved with the following parameters:

- $I = 0.95$ ;  $n = 150$ ;  $h = 1000$

The best fit of the Yule-Simon Model with Memory has been achieved with the following parameters:

- $p = 0.102$ ;  $\tau = 480$ ;  $n_0 = 100$

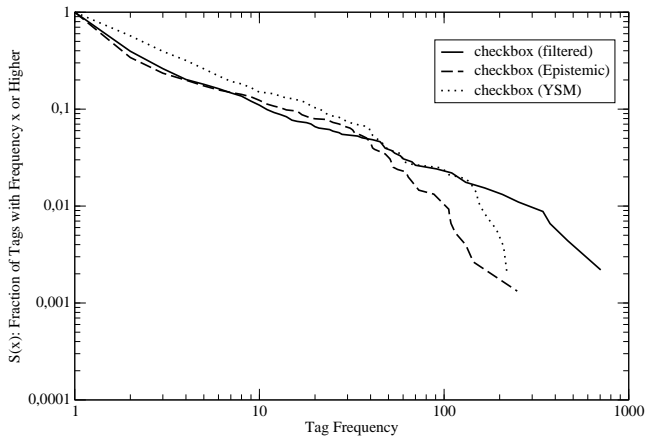


Figure 41: Empirical distribution functions for the filtered *checkbox* and the best fitting graph simulated with the Epistemic Model and the Yule-Simon Model with Memory (YSM).

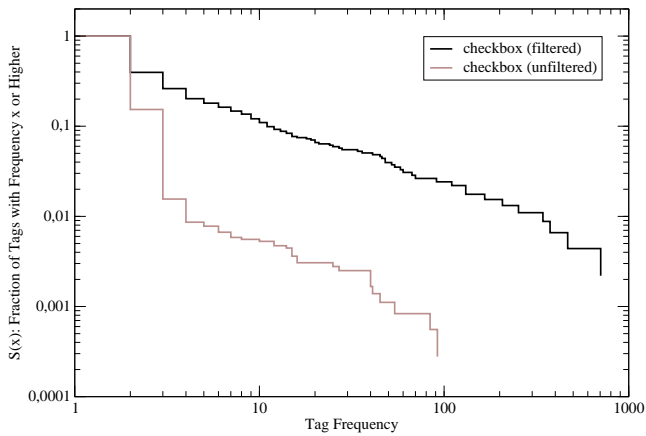


Figure 42: Empirical distribution functions for the *checkbox* stream pair.

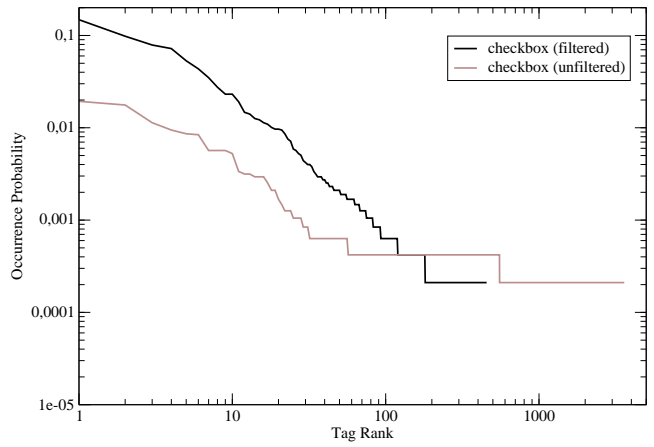


Figure 43: Frequency-rank plot for the *checkbox* stream pair.

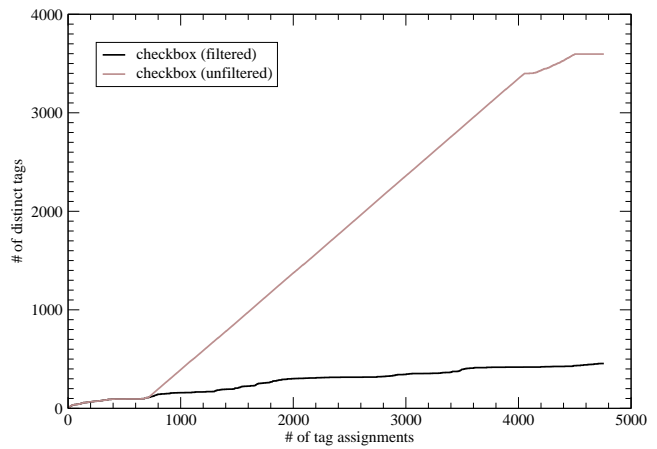


Figure 44: Vocabulary growth for the *checkbox* stream pair.

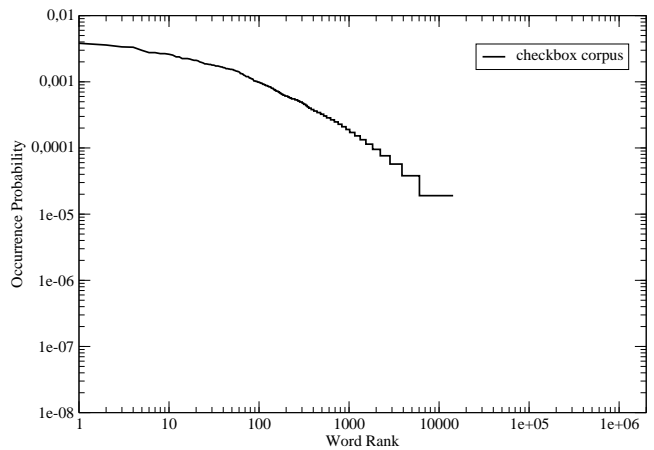


Figure 45: Frequency-rank plot for the *checkbox* web corpus.

## 12. DATAWAREHOUSE STREAM PAIR

For the filtered datawarehouse stream, the best fit of the Epistemic Model has been achieved with the following parameters:

- $I = 0.95$ ;  $n = 250$ ;  $h = 3000$

The best fit of the Yule-Simon Model with Memory has been achieved with the following parameters:

- $p = 0.191$ ;  $\tau = 500$ ;  $n_0 = 100$

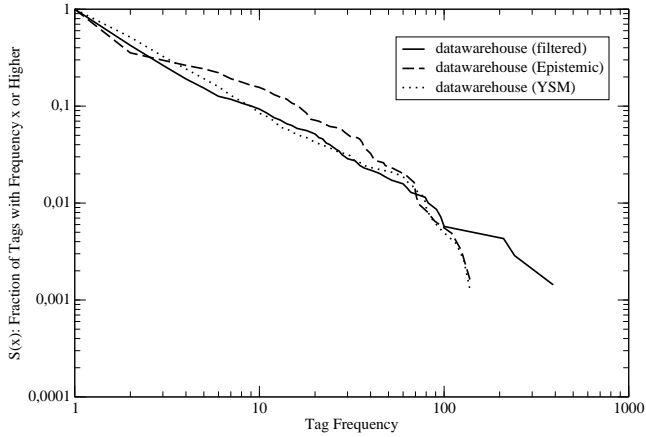


Figure 46: Empirical distribution functions for the filtered *datawarehouse* and the best fitting graph simulated with the Epistemic Model and the Yule-Simon Model with Memory (YSM).

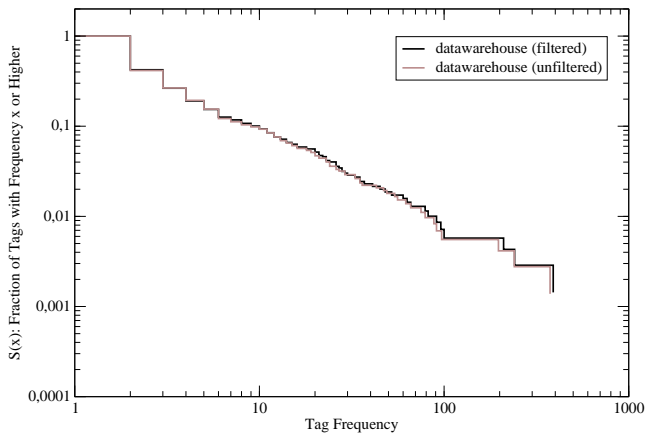


Figure 47: Empirical distribution functions for the *datawarehouse* stream pair.

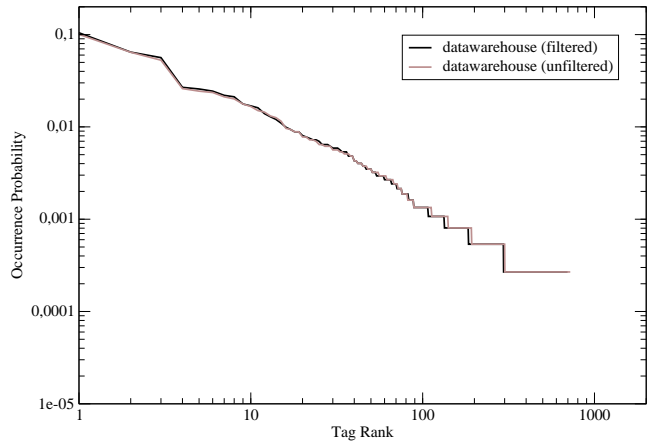


Figure 48: Frequency-rank plot for the *datawarehouse* stream pair.

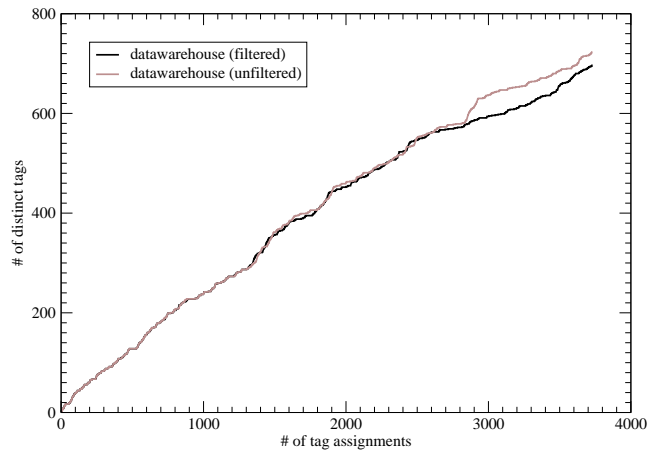


Figure 49: Vocabulary growth for the *datawarehouse* stream pair.

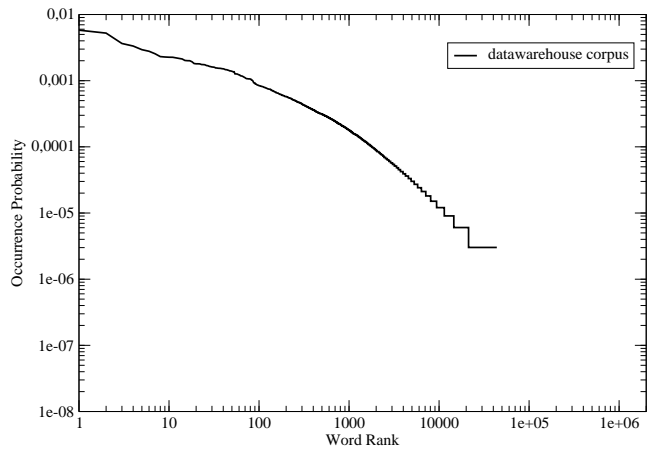


Figure 50: Frequency-rank plot for the *datawarehouse* web corpus.

### 13. TOOLS STREAM PAIR

For the filtered tools stream, the best fit of the Epistemic Model has been achieved with the following parameters:

- $I = 0.947$ ;  $n = 1000$ ;  $h = 7000$

The best fit of the Yule-Simon Model with Memory has been achieved with the following parameters:

- $p = 0.144$ ;  $\tau = 500$ ;  $n_0 = 100$

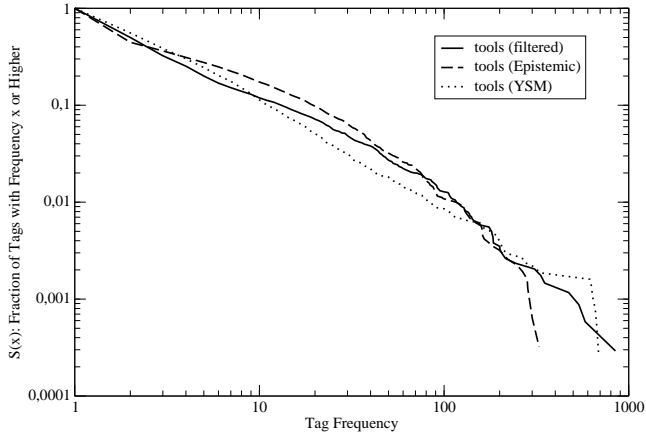


Figure 51: Empirical distribution functions for the filtered *tools* and the best fitting graph simulated with the Epistemic Model and the Yule-Simon Model with Memory (YSM).

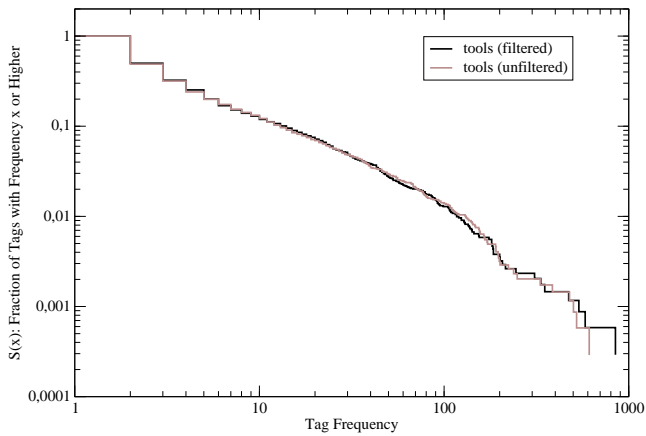


Figure 52: Empirical distribution functions for the *tools* stream pair.

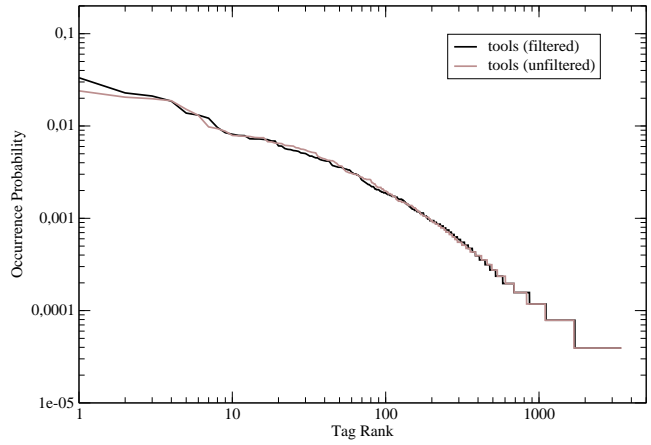


Figure 53: Frequency-rank plot for the *tools* stream pair.

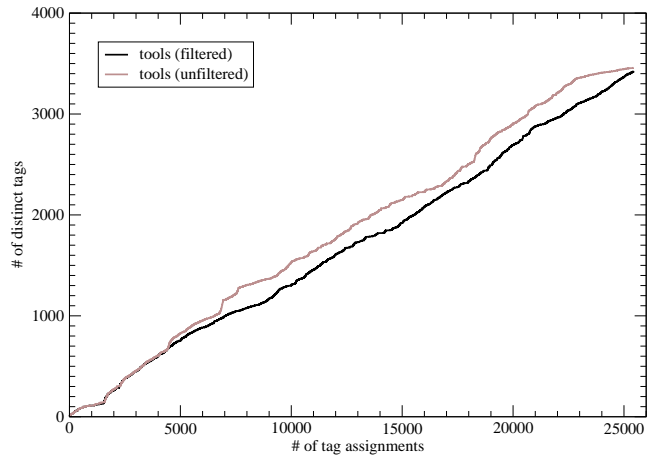


Figure 54: Vocabulary growth for the *tools* stream pair.

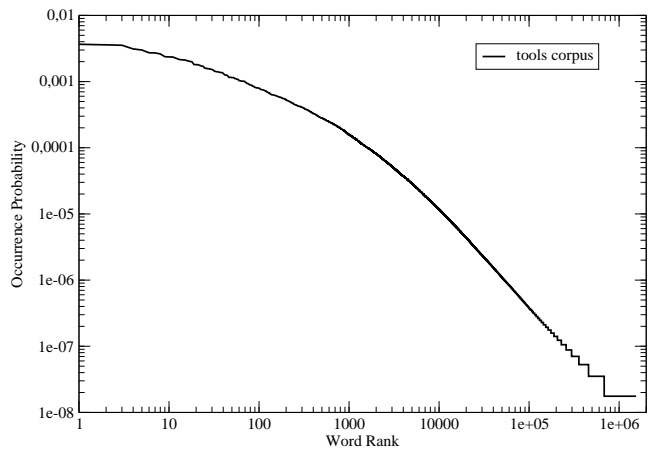


Figure 55: Frequency-rank plot for the *tools* web corpus.

## 14. SOCIAL STREAM PAIR

For the filtered social stream, the best fit of the Epistemic Model has been achieved with the following parameters:

- $I = 0.763$ ;  $n = 1350$ ;  $h = 5000$

The best fit of the Yule-Simon Model with Memory has been achieved with the following parameters:

- $p = 0.24$ ;  $\tau = 500$ ;  $n_0 = 100$

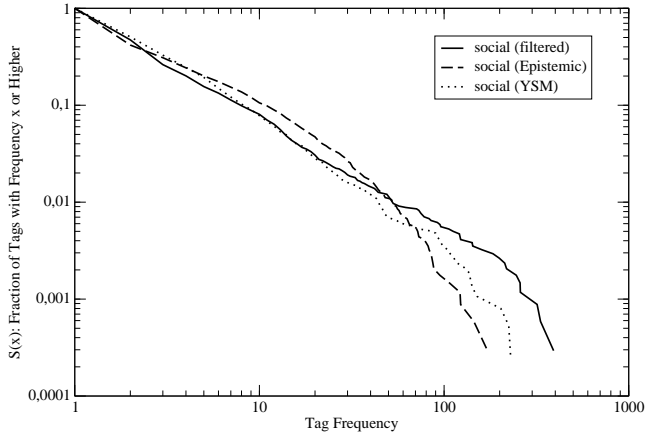


Figure 56: Empirical distribution functions for the filtered *social* and the best fitting graph simulated with the Epistemic Model and the Yule-Simon Model with Memory (YSM).

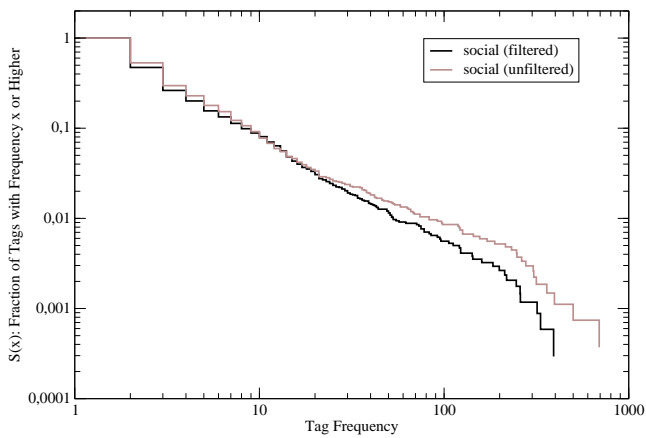


Figure 57: Empirical distribution functions for the *social* stream pair.

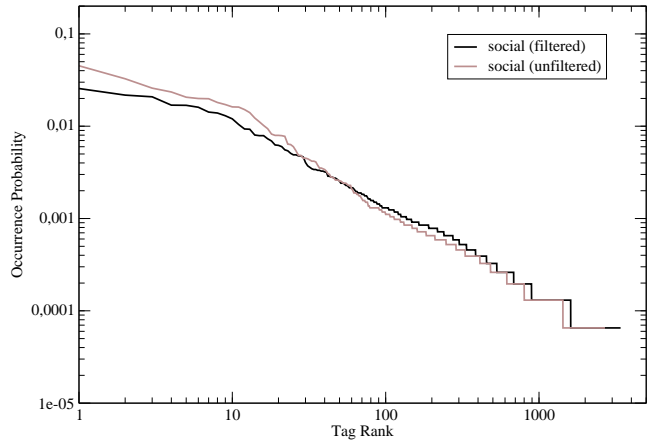


Figure 58: Frequency-rank plot for the *social* stream pair.

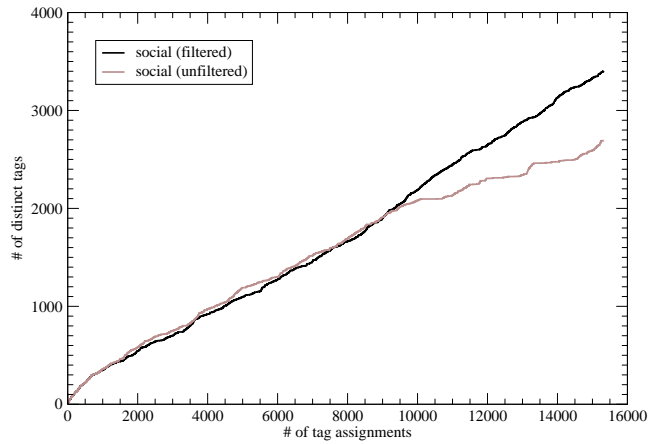


Figure 59: Vocabulary growth for the *social* stream pair.

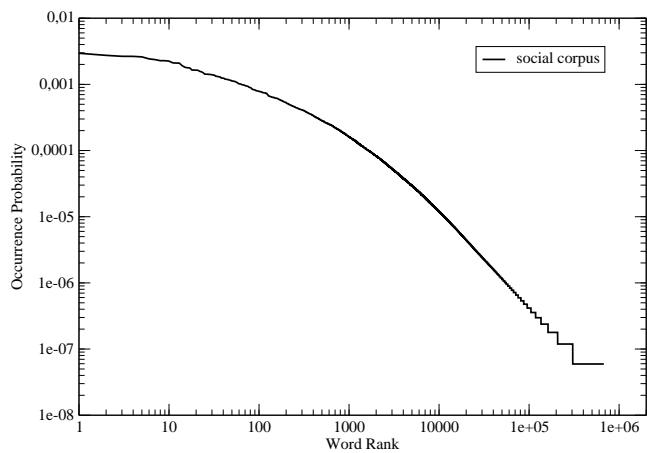


Figure 60: Frequency-rank plot for the *social* web corpus.

## 15. DESIGN STREAM PAIR

For the filtered design stream, the best fit of the Epistemic Model has been achieved with the following parameters:

- $I = 0.876$ ;  $n = 600$ ;  $h = 3000$

The best fit of the Yule-Simon Model with Memory has been achieved with the following parameters:

- $p = 0.204$ ;  $\tau = 500$ ;  $n_0 = 100$

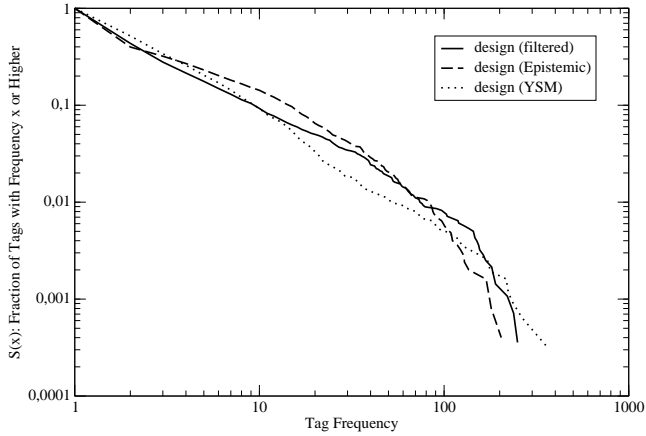


Figure 61: Empirical distribution functions for the filtered *design* and the best fitting graph simulated with the Epistemic Model and the Yule-Simon Model with Memory (YSM).

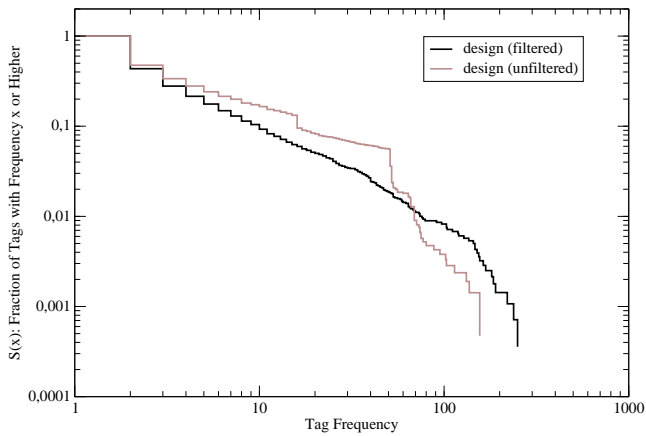


Figure 62: Empirical distribution functions for the *design* stream pair.

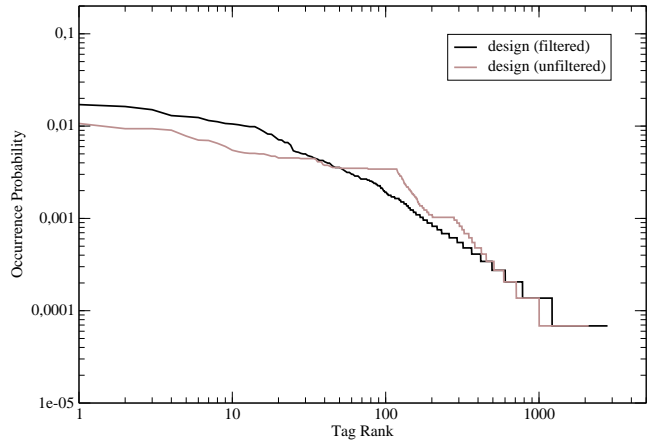


Figure 63: Frequency-rank plot for the *design* stream pair.

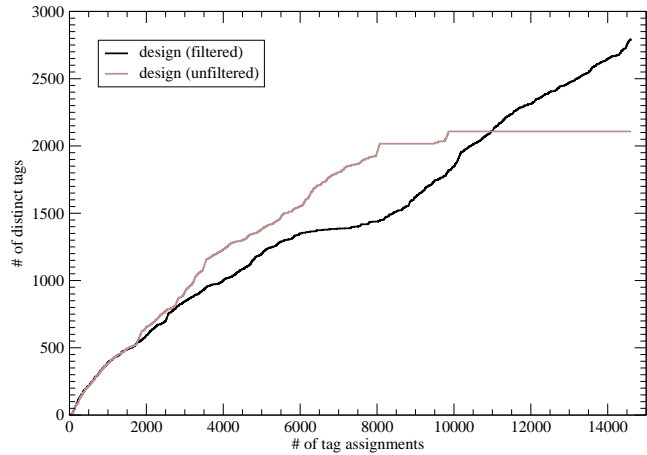


Figure 64: Vocabulary growth for the *design* stream pair.

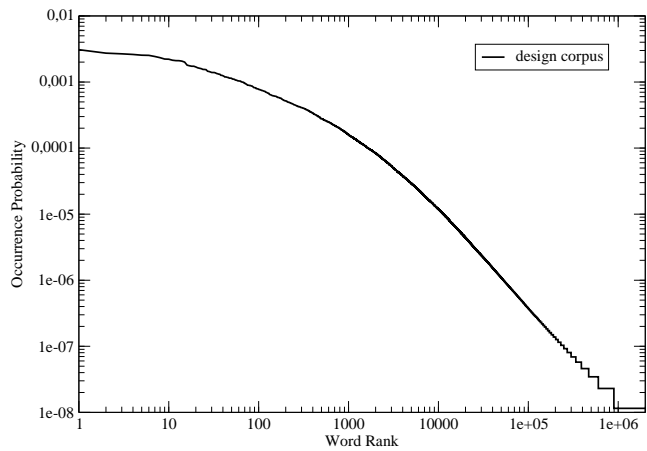


Figure 65: Frequency-rank plot for the *design* web corpus.

## 16. ANALYSIS STREAM PAIR

For the filtered analysis stream, the best fit of the Epistemic Model has been achieved with the following parameters:

- $I = 0.95$ ;  $n = 400$ ;  $h = 5000$

The best fit of the Yule-Simon Model with Memory has been achieved with the following parameters:

- $p = 0.104$ ;  $\tau = 500$ ;  $n_0 = 100$

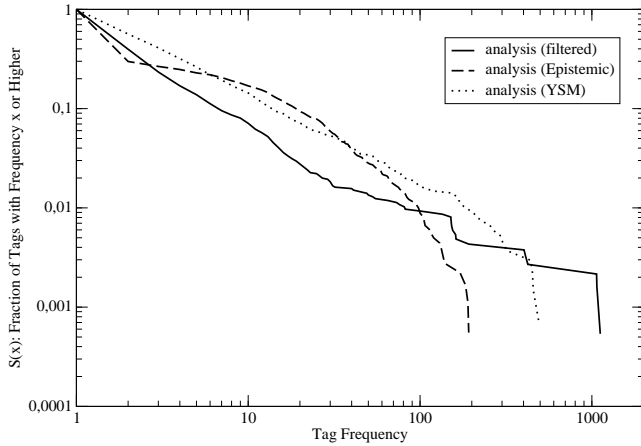


Figure 66: Empirical distribution functions for the filtered *analysis* and the best fitting graph simulated with the Epistemic Model and the Yule-Simon Model with Memory (YSM).

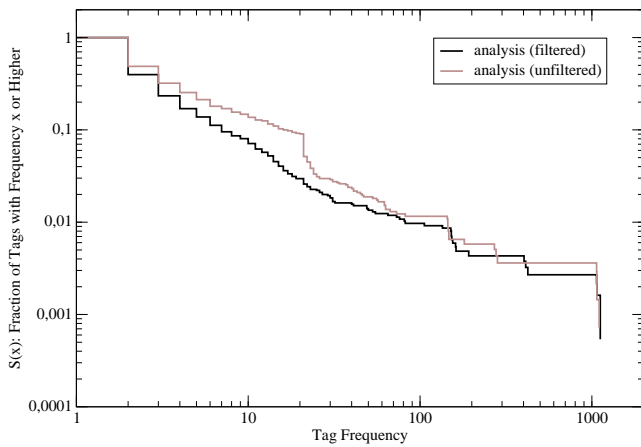


Figure 67: Empirical distribution functions for the *analysis* stream pair.

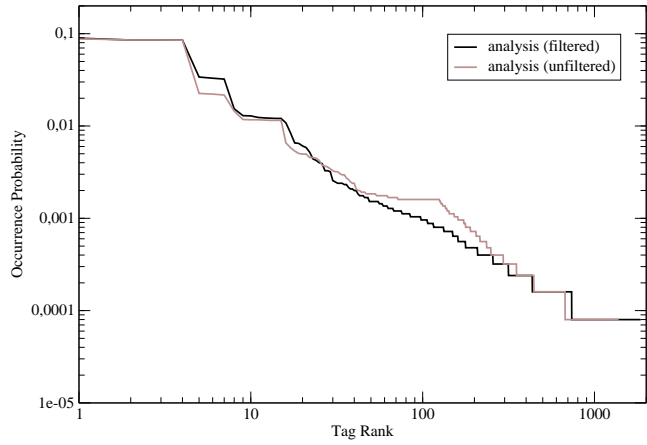


Figure 68: Frequency-rank plot for the *analysis* stream pair.

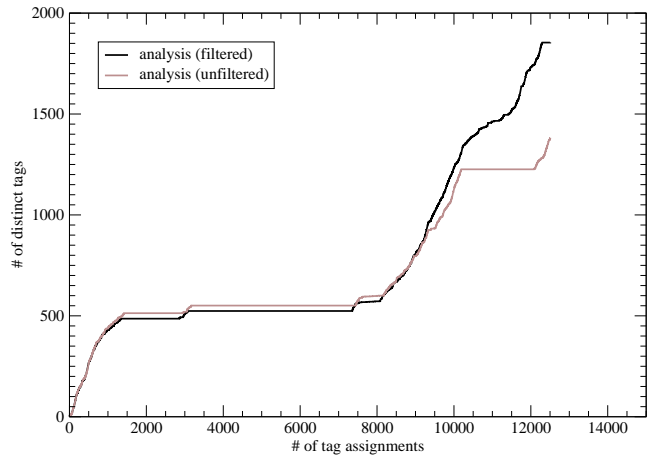


Figure 69: Vocabulary growth for the *analysis* stream pair.

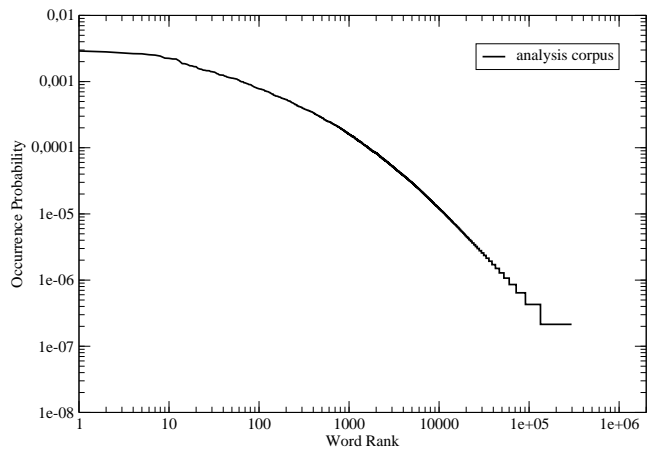


Figure 70: Frequency-rank plot for the *analysis* web corpus.

## 17. BLOGS STREAM PAIR

For the filtered blogs stream, the best fit of the Epistemic Model has been achieved with the following parameters:

- $I = 0.782$ ;  $n = 700$ ;  $h = 3000$

The best fit of the Yule-Simon Model with Memory has been achieved with the following parameters:

- $p = 0.265$ ;  $\tau = 500$ ;  $n_0 = 100$

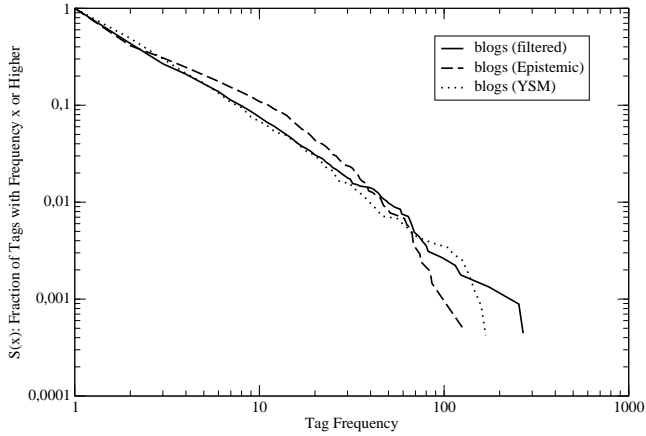


Figure 71: Empirical distribution functions for the filtered *blogs* and the best fitting graph simulated with the Epistemic Model and the Yule-Simon Model with Memory (YSM).

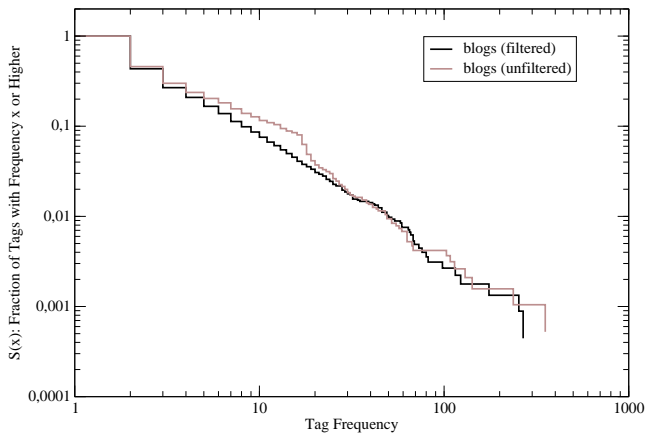


Figure 72: Empirical distribution functions for the *blogs* stream pair.

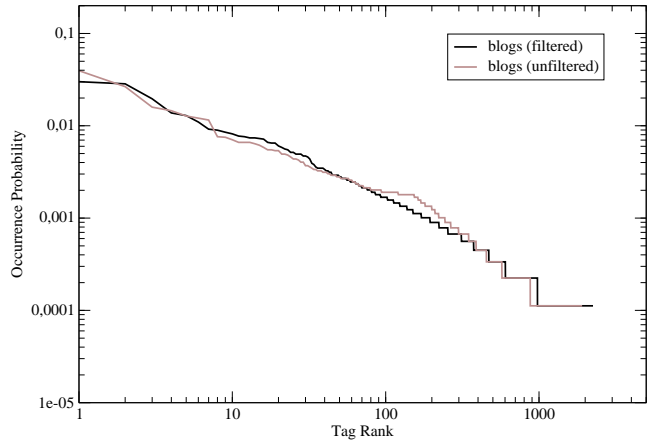


Figure 73: Frequency-rank plot for the *blogs* stream pair.

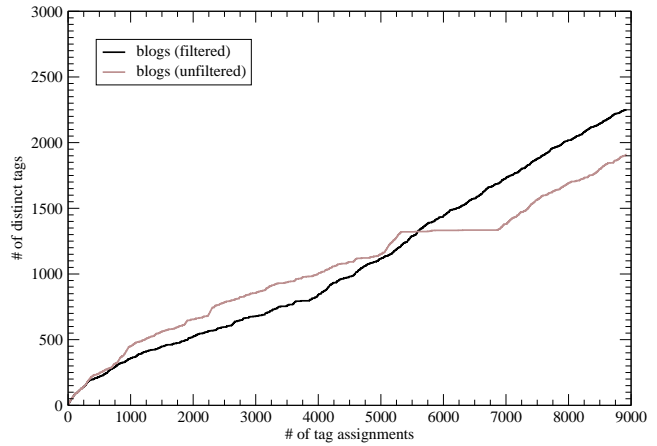


Figure 74: Vocabulary growth for the *blogs* stream pair.

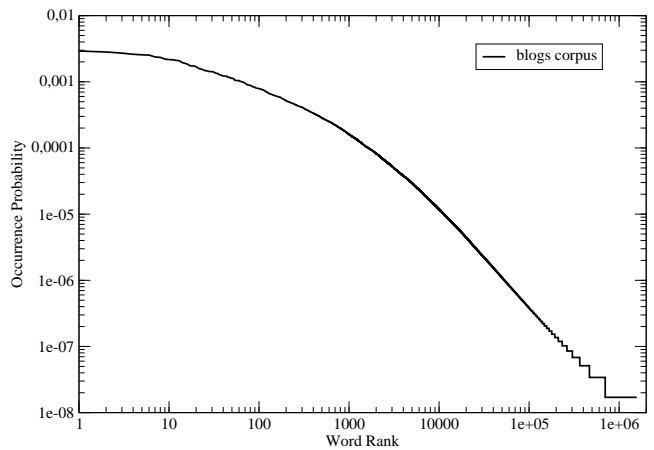


Figure 75: Frequency-rank plot for the *blogs* web corpus.