Precursors and Laggards

An Analysis of Semantic Temporal Relationships on a Blog Network

Telmo Menezes, Camille Roth & Jean-Philippe Cointet

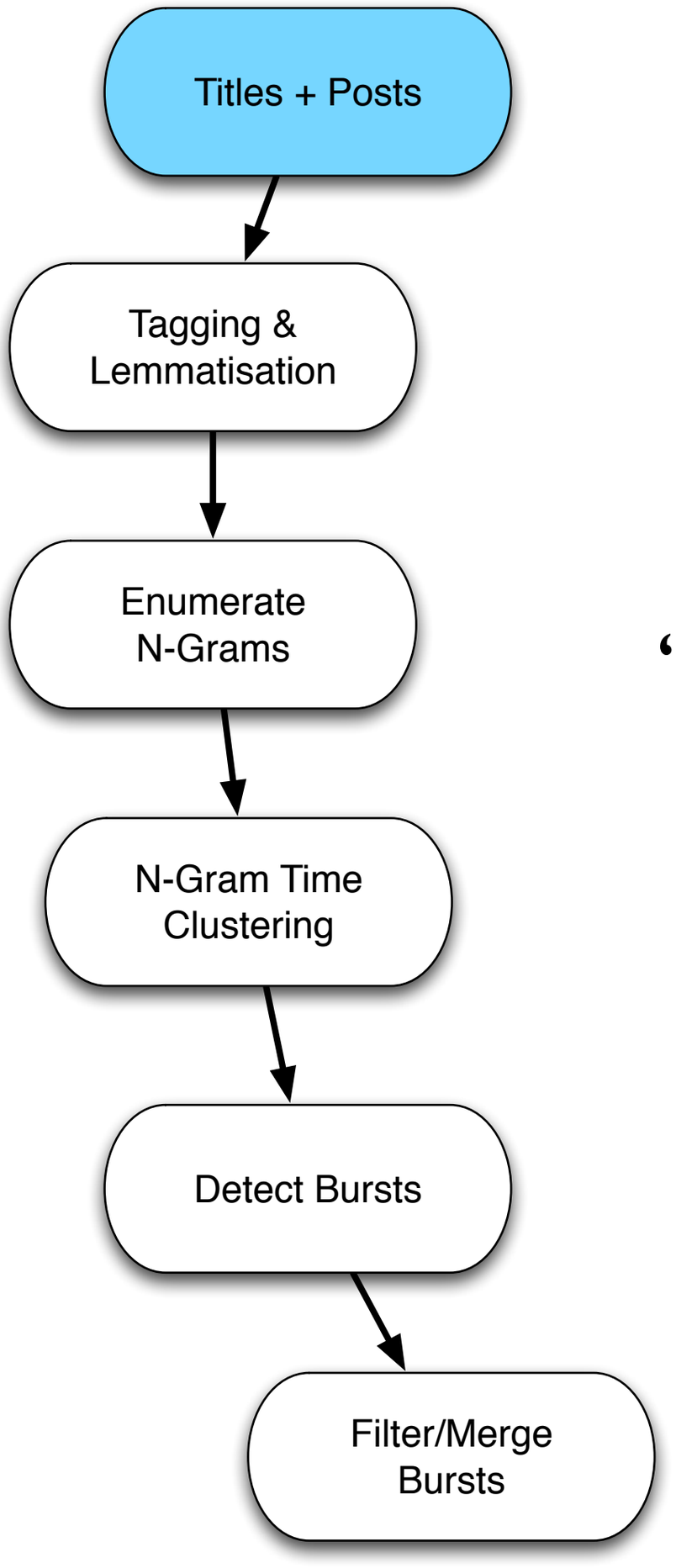
CNRS - French National Center For Scientific Research
ISC-PIF - Institut des Systèmes Complexes - Paris Île-de-France
CREA - Centre de Recherche en Épistémologie Appliquée

Goals

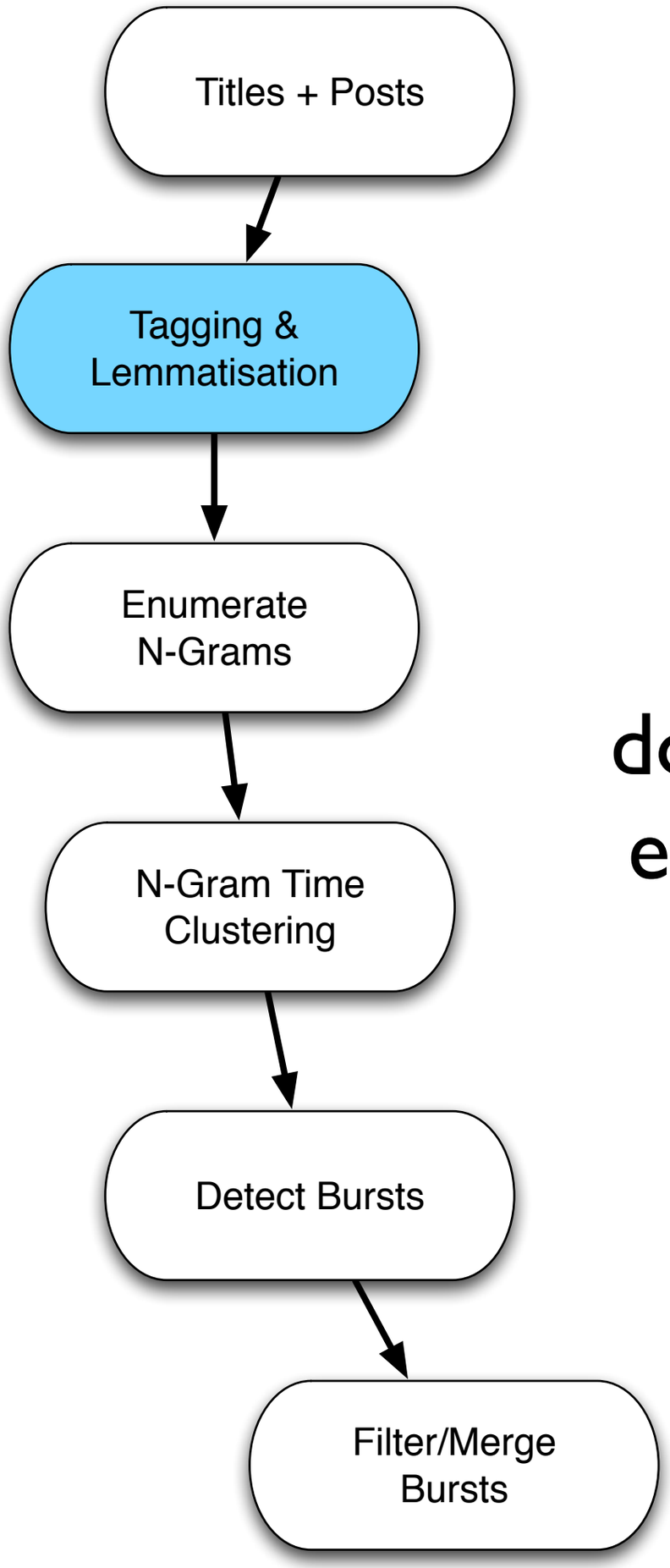
- Quantify the temporal relationships between blogs at the semantic level;
- Show that these metrics adds to the knowledge obtainable by considering only the structural level.

Overview

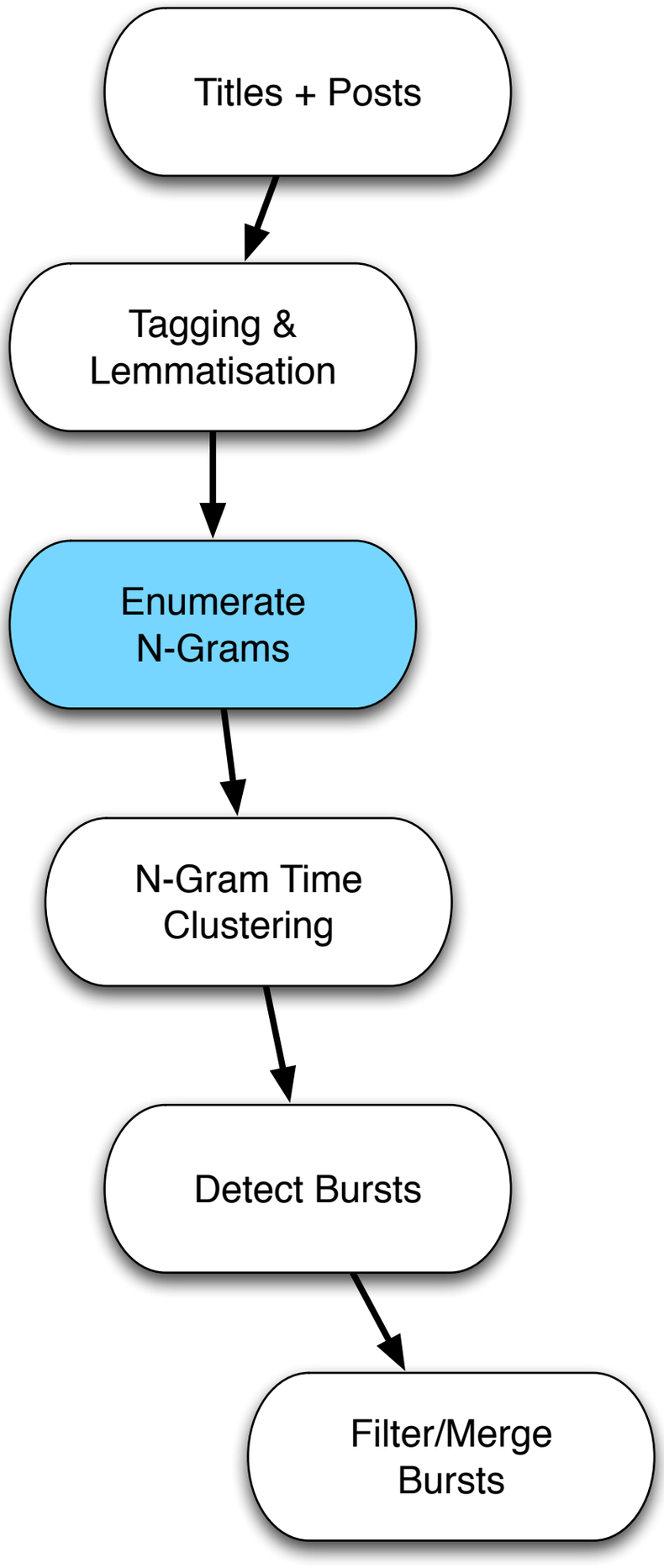
- Semantic "Unit of Activity" Detection
- Probabilistic Precursor/Laggard Scoring
- Results (French political blogosphere)



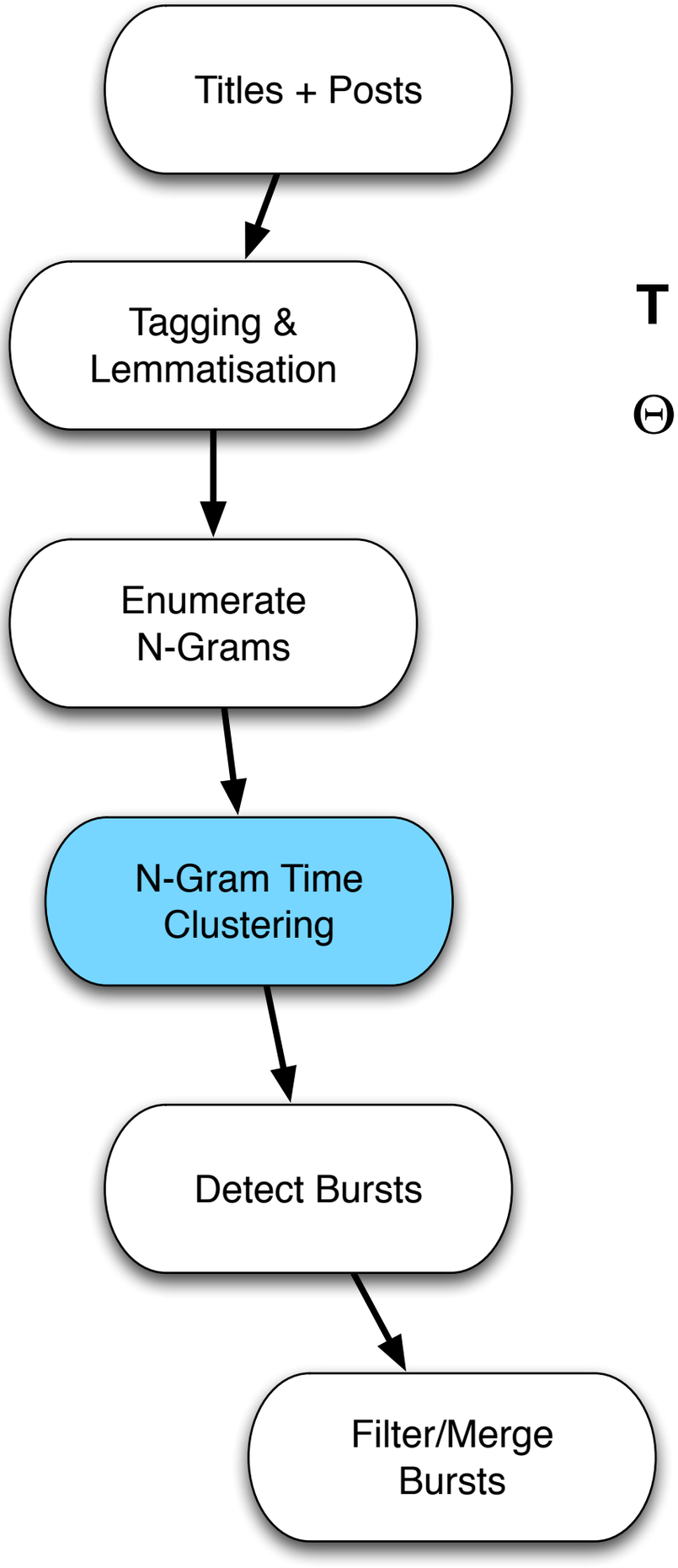
“[...] Rencontre entre Dominique de Villepin et des jeunes [...]”

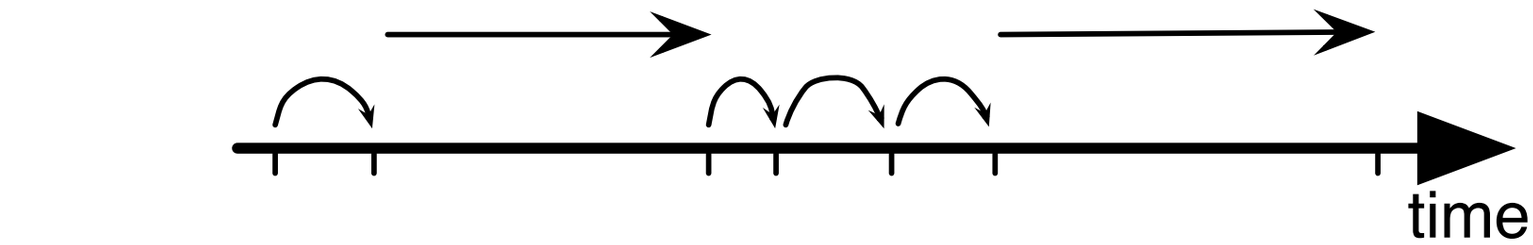


“[...] rencontre/NOM entre/PRP dominique/NAM de/PRP villepin/NAM et/KON du/PRP:det jeune/NOM [...]”



- Two or more words per n-gram
- At least one noun
- Reject words that are not nouns, verbs, adjectives or numbers
- Reject words with strong chronological meanings (Saturday, April, Spring, Christmas, etc..)



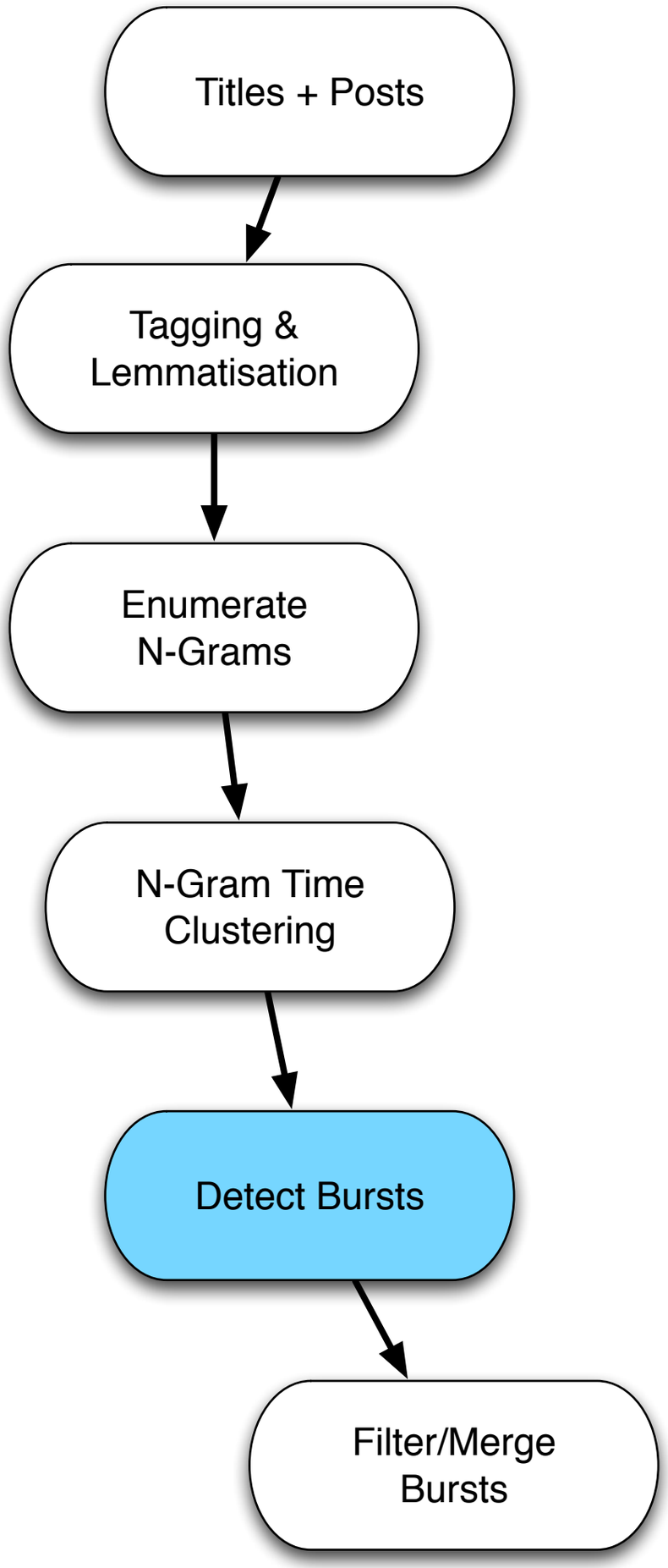


T	t_1	t_2	t_3	t_4	t_5	t_6	t_7
Θ	0	1	0	0	0	1	1

$$V_{\longmapsto}(T,\Theta) = \frac{\sum_{i=1}^{|T|-1}(t_{i+1}-t_i)\theta_i}{\sum_{i=1}^{|T|-1}\theta_i},$$
$$\text{if } \sum_{i=1}^{|T|-1}\theta_i > 0, 0 \text{ otherwise}$$

$$V_{\curvearrowright}(T,\Theta) = \frac{\sum_{i=1}^{|T|-1}(t_{i+1}-t_i)(1-\theta_i)}{\sum_{i=1}^{|T|-1}(1-\theta_i)},$$
$$\text{if } \sum_{i=1}^{|T|-1}\theta_i > 0, 0 \text{ otherwise}$$

$$\rho(T,\Theta) = \frac{V_{\longmapsto}(T,\Theta)}{V_{\curvearrowright}(T,\Theta)}, \text{if } V_{\curvearrowright}(T,\Theta) > 0 \text{ , } 0 \text{ otherwise}$$



- Minimum number of participating blogs (4)
- Minimum average time between posts (1 hour)
- Maximum average time between posts (3 days)
- Minimum burst duration (3 days)
- Maximum total duration of bursts with same n-gram (1 month)

Titles + Posts

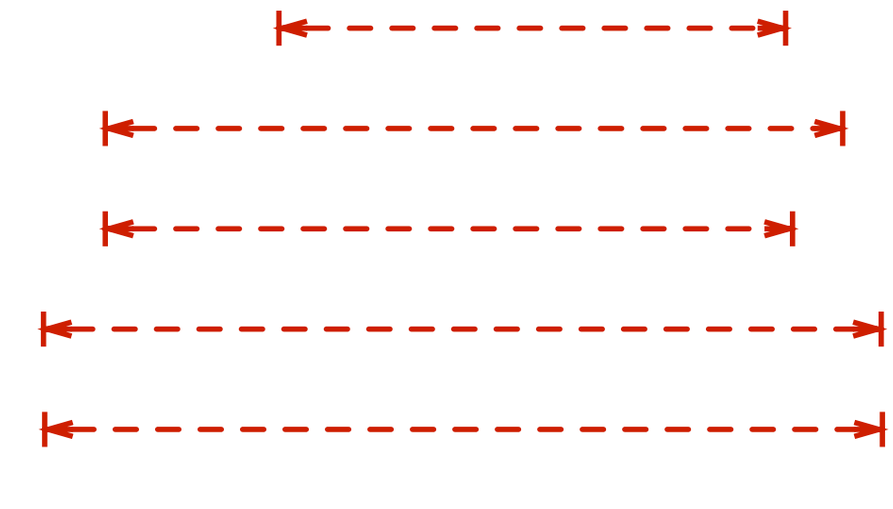
Tagging & Lemmatisation

Enumerate N-Grams

N-Gram Time Clustering

Detect Bursts

Filter/Merge Bursts



région avoir apporter contribution débat

apporter contribution débat

contribution débat national

contribution débat identité nationale

contribution débat identité

Probabilistic Scoring

- Given two blogs, what's the relationship of temporal precedence of one over the other, discounting asymmetrical posting rates?

Probabilistic Scoring

A = {topics where both blogs participate}
Y = {topics where b participates before b'}
C = vector of the probabilities of b participating in a topic before b' by chance
Z = {topics where b is hypothesized to display a behavior of precedence over b'}
R = {topics where b is hypothesized to have preceded b' by chance}

$$\lambda(\gamma(b,b') = p|A,Y,C) =$$

$$\sum_{\substack{Z \cup R = Y \\ Z \cap R = \emptyset}} \lambda(\gamma(b,b') = p|A,Y,C,Z,R)$$

Probabilistic Scoring

$$\lambda(\gamma(b, b') = p|A, Y, C, Z, R) = P_Z(A, Z, p) \cdot P_R(A, R, C)$$

$$P_Z(A, Z, p) = p^{|Z|}(1 - p)^{|A| - |Z|}$$

$$P_R(A, R, C) = \prod_{r \in R} C_r \prod_{r \in A \setminus R} 1 - C_r$$

$$C_r = \frac{Np(b, [t_s(r); t_e(r)])}{Np(b, [t_s(r); t_e(r)]) + Np(b', [t_s(r); t_e(r)])}$$

Dyadic Precursor Score

$$\gamma(b, b') = \frac{\int_0^1 l(\gamma(b, b') = p|A, Y, C) \cdot p \cdot dp}{\int_0^1 l(\gamma(b, b') = p|A, Y, C) \cdot dp}$$

Adjusted Dyadic Precursor Score

M = the event of b participating before b' under a precedence relationship
H = the event of b and b' participating on a same topic

$$\gamma(b, b') = P_r(M|H)$$

$$P_r(M|H) = \frac{P_r(H|M)P_r(M)}{P_r(H)}$$

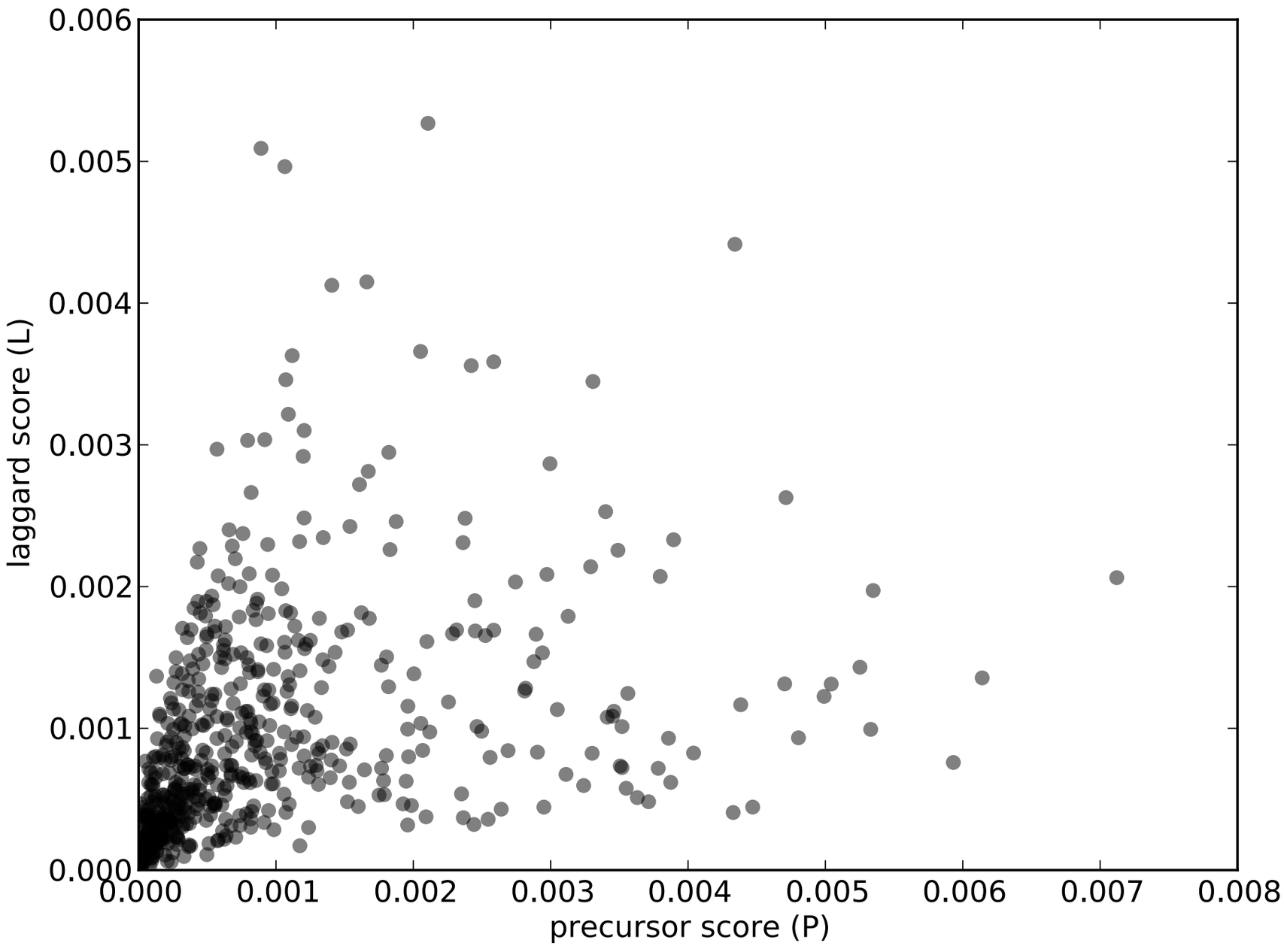
$$\omega(b, b') = P_r(M) = P_r(M|H)P_r(H) = \gamma(b, b')P_r(H)$$

Global Precursor and Laggard Scores

$$P(b) = \frac{1}{|B| - 1} \sum_{b' \in B \setminus \{b\}} \omega(b, b')$$

$$L(b) = \frac{1}{|B| - 1} \sum_{b' \in B \setminus \{b\}} \omega(b', b)$$

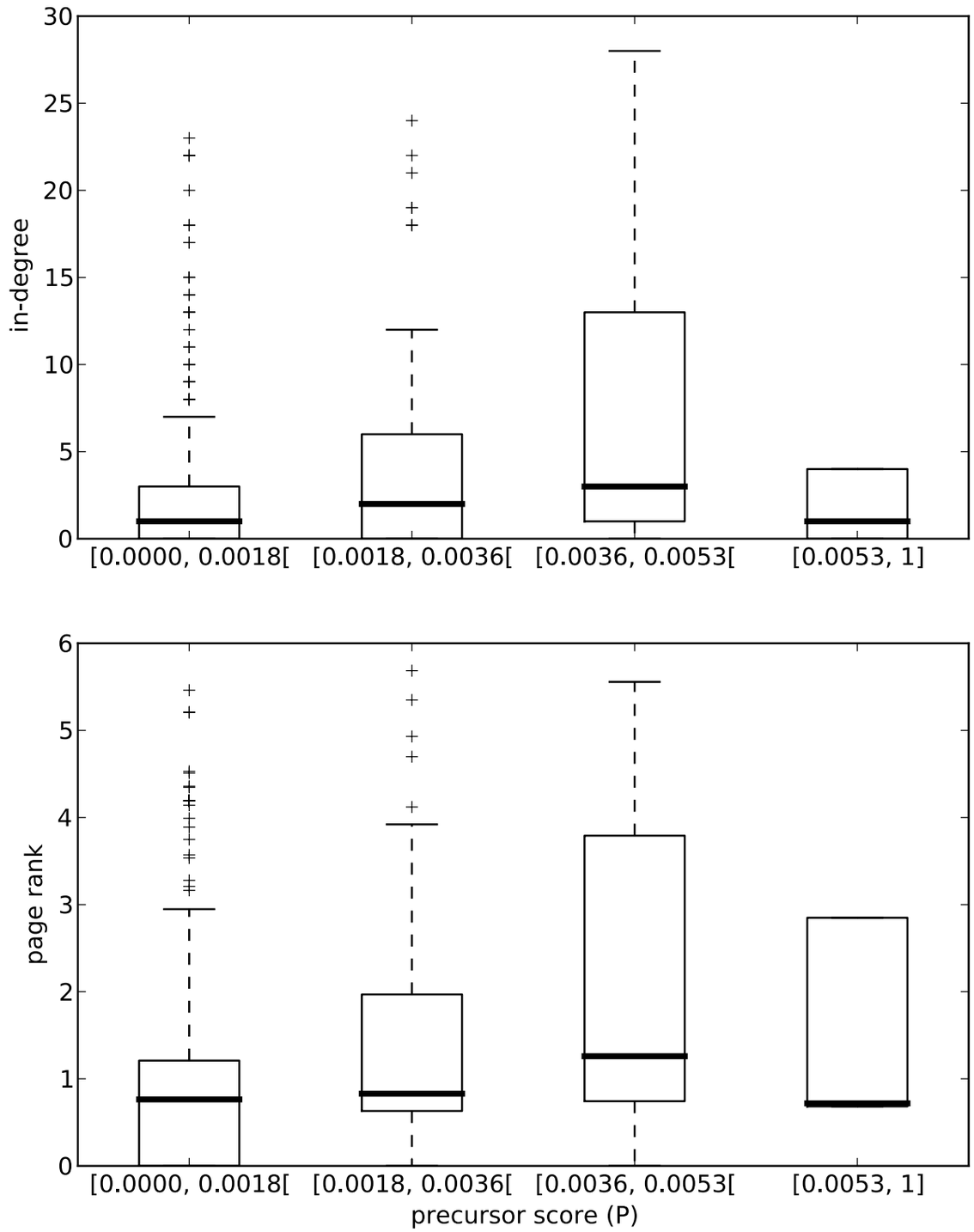
Precursor vs Laggard Scores



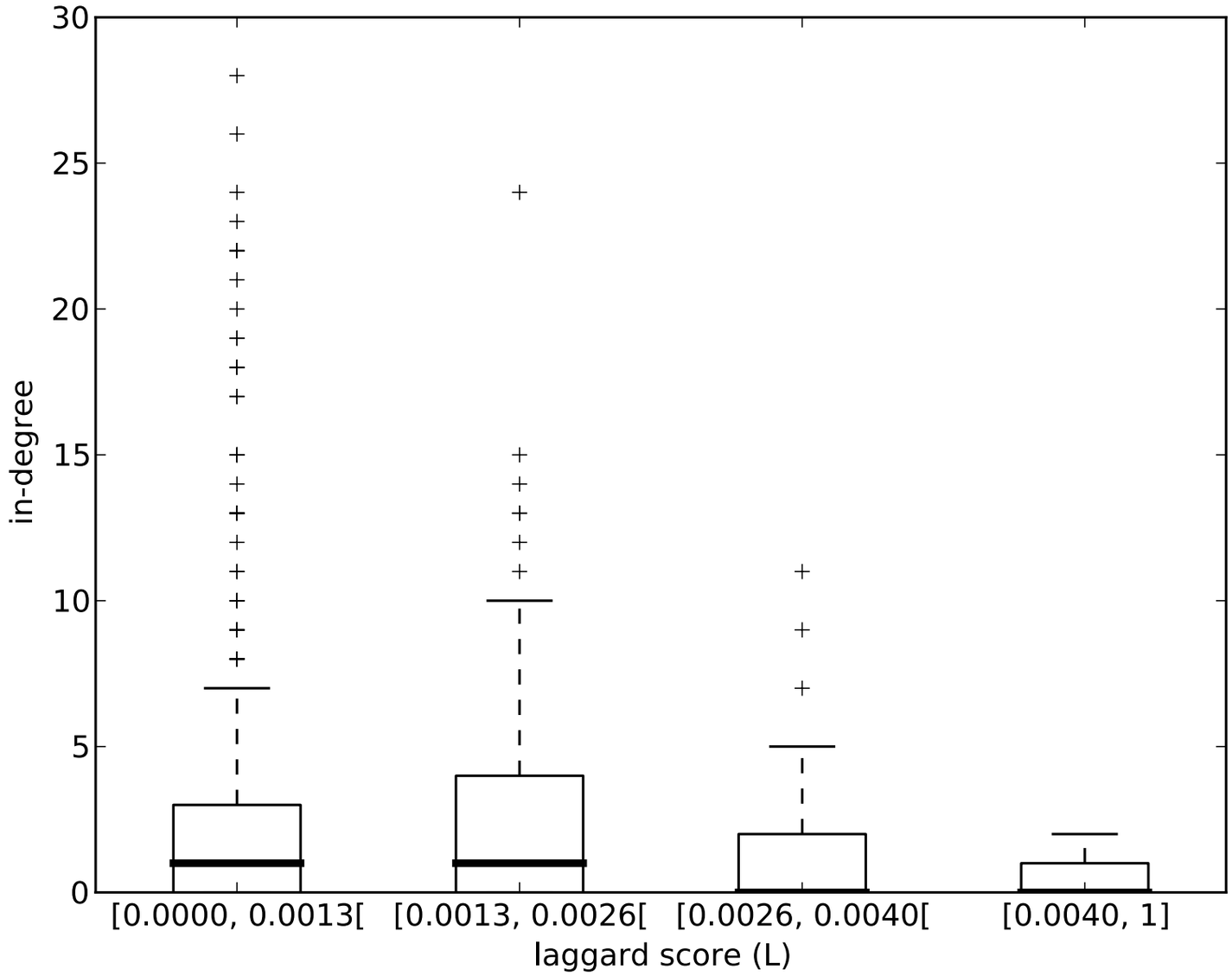
Significance of Mean In-Degree Relationships

		2.08	6.19	1.59	3.50
		p/	**P**/	*p***L**	**PL**
2.08	*p*/				
6.19	**P**/	**			
1.59	*p***L**	*	***		
3.50	**PL**			***	

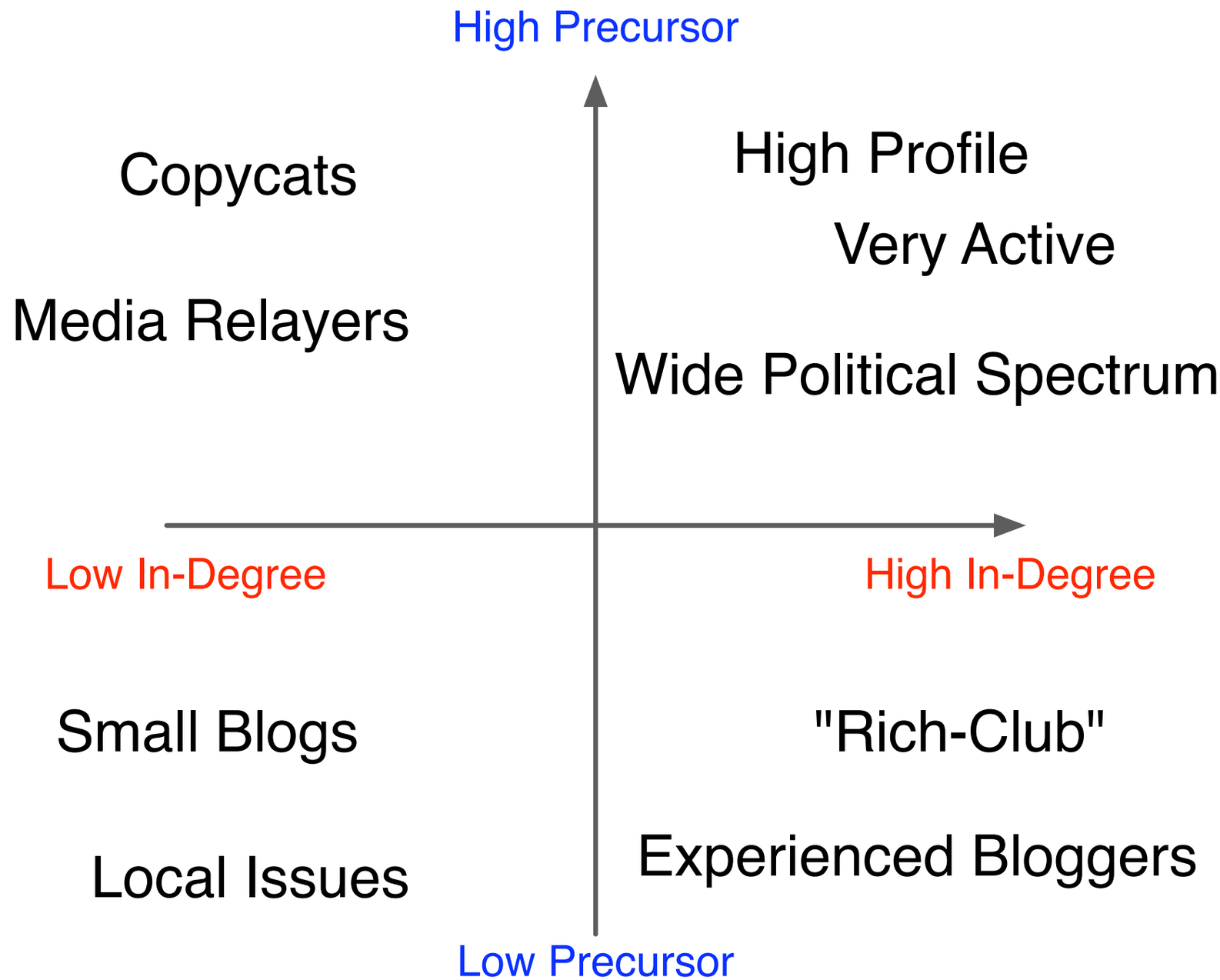
In-Degree / Page Rank Distribution per Precursor Score Interval



In-Degree Distribution per Laggard Score Interval



2D Clustering



Conclusions

- Detection of semantic units of iteration
- Dyadic and global precursor/laggard scores
- Fully automated process
- Scores add to the information obtainable by structural metrics
- Method validated by a blind test with an expert