You are the way you (structurally) talk: Structural-temporal neighbourhoods of posts to characterize users in online forums

Alberto Lumbreras
Jouve B., Velcin J., Guégan, M.

May 3, 2016

Overview

The data

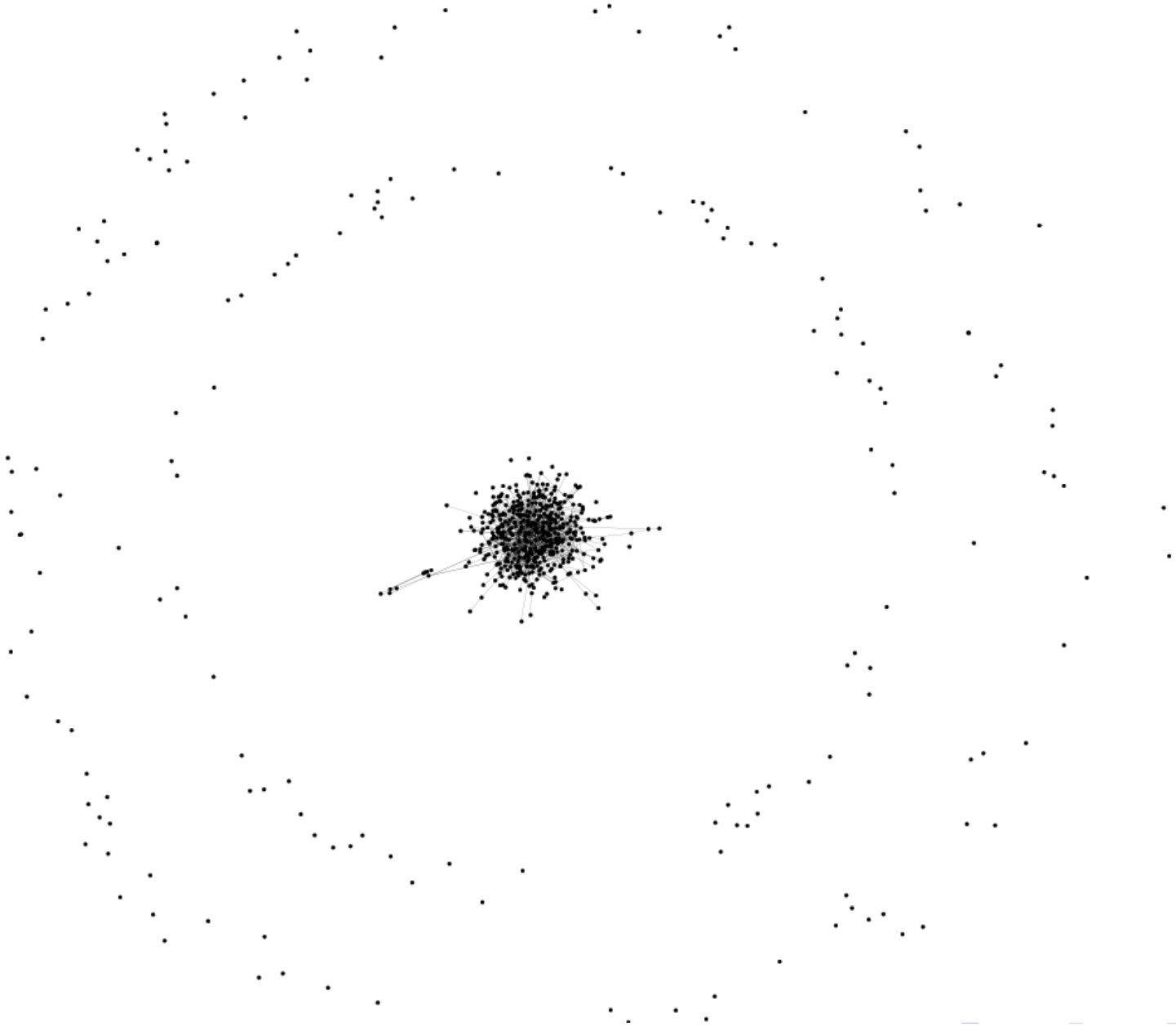
Reddit. A forum of forums

Download monthly dumps from:
http://couch.whatbox.ca:36975/reddit/comments/monthly/

Extract forum of interest:
www.reddit.com/r/science
www.reddit.com/r/france
www.reddit.com/r/sociology
www.reddit.com/r/complexsystems
www.reddit.com/r/gameofthrones ← in this presentation
www.reddit.com/r/podemos ← in this presentation
...

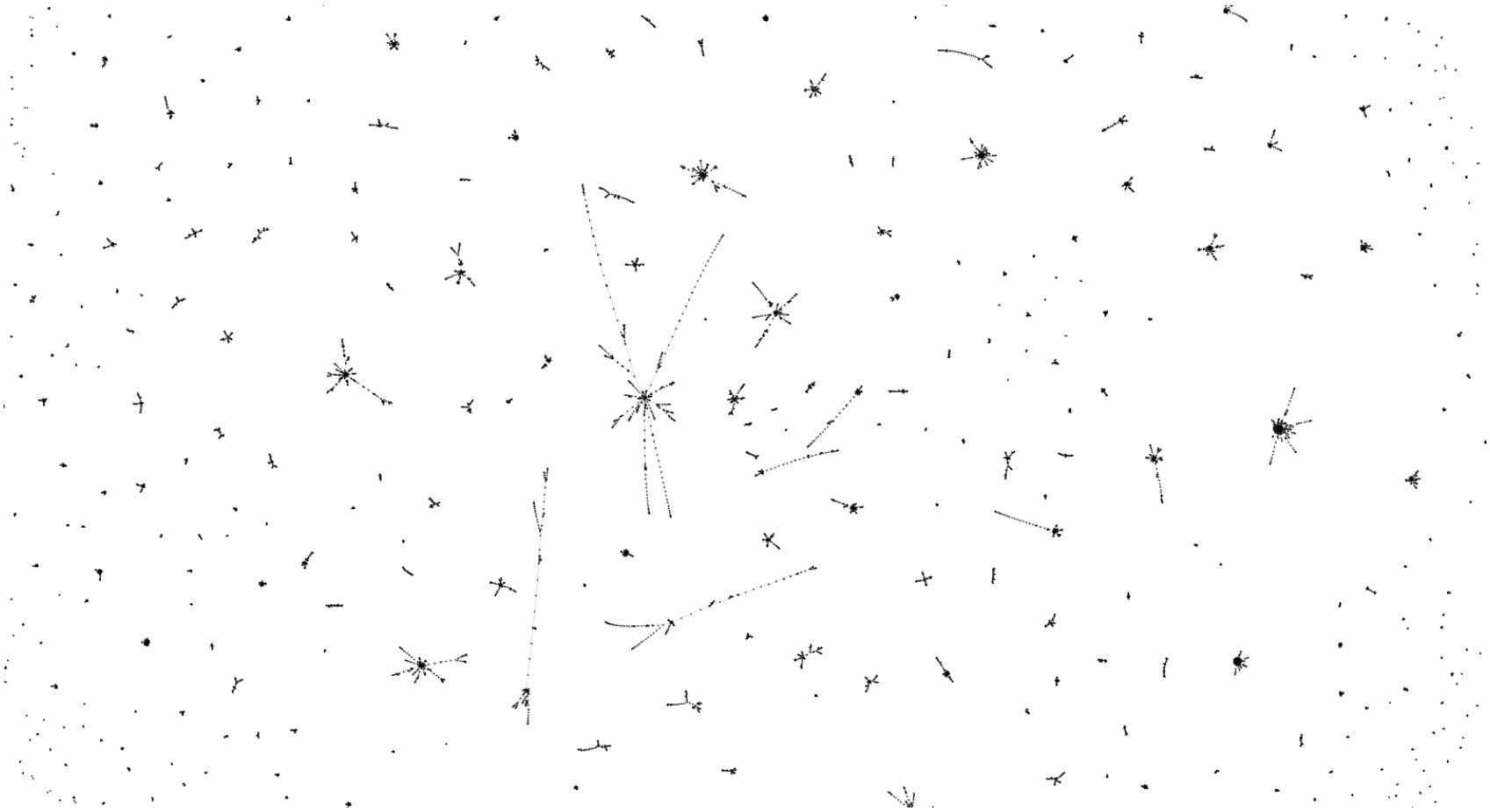
Graph representations

Graph of user interactions (a social network)



Graph representations

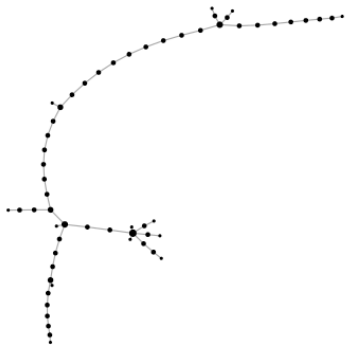
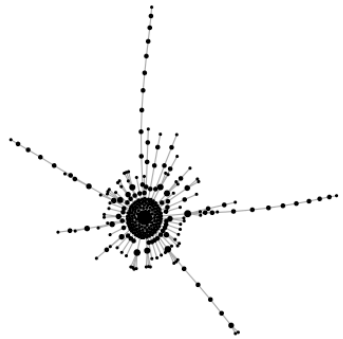
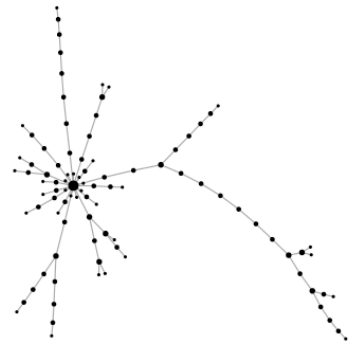
Trees of posts



Graph representations

Conversations are trees

- Explicit structure.
- Dynamic (order, time)

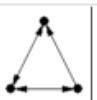
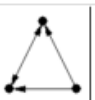
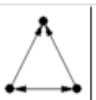




Intuition

Hypothesis: different individuals have tendency towards different types of conversations and these types are reflected in the structure of their interactions.

Triadic structures

Triads are not enough

Motif															
Motif ID			36	164	12	14	6	78	38	174	166	46	238	102	140

Triads in **trees of posts**:

- Only 3 possible triads (dyad, chain and star)

Triads in **social graph**:

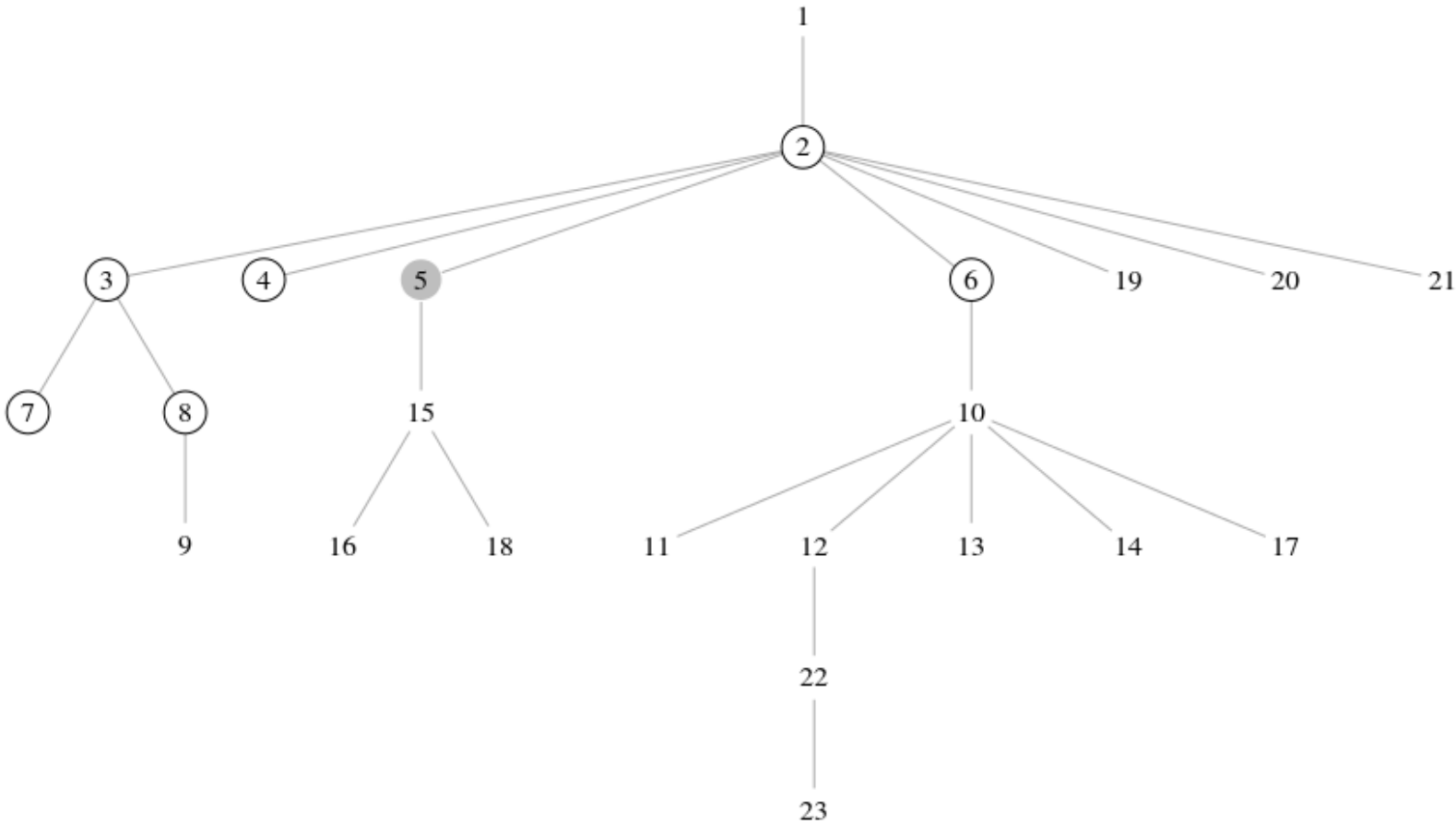
- Order (therefore dynamic) is missing.

We need something richer that captures the dynamics of conversations.

Order-based neighbourhoods

Definition

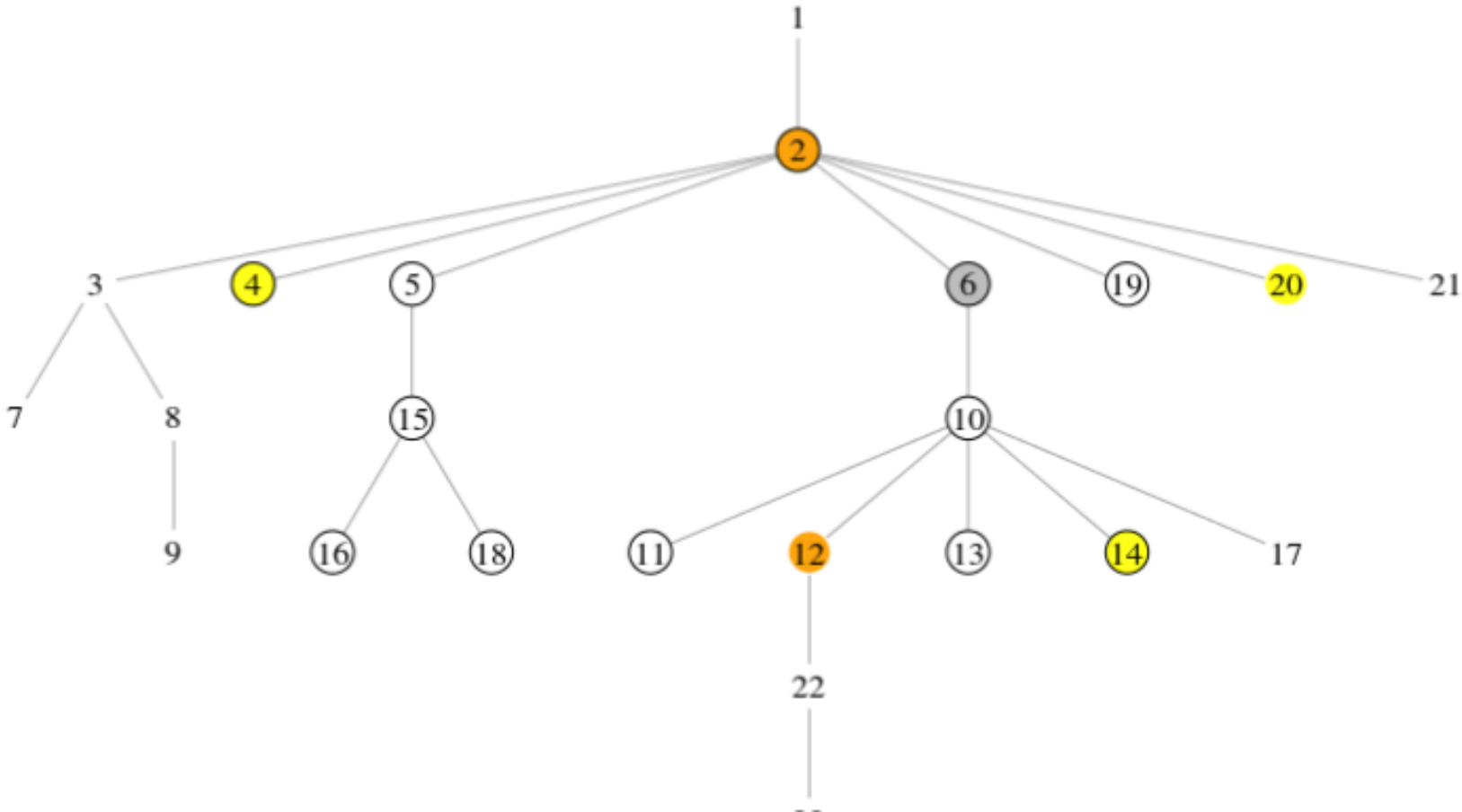
- 1. Extract neighbourhood of post i with radius r.
- 2. Keep only the n posts that are closest (in time) to post i.



Time-based neighbourhoods

Definition

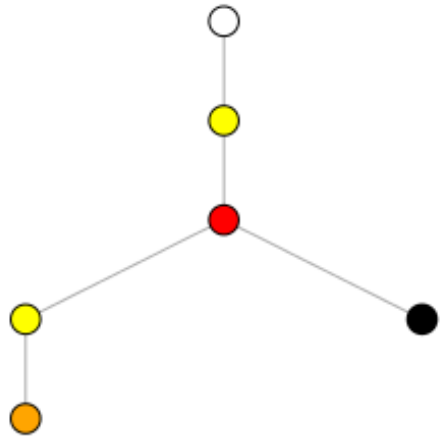
- 1. Extract neighbourhood of post i with radius r.
- 2. Detect changes of speed (vertical/horizontal *changepoints*) (PELT algo)
- 3. From i, get the posts around until a changepoint is found.



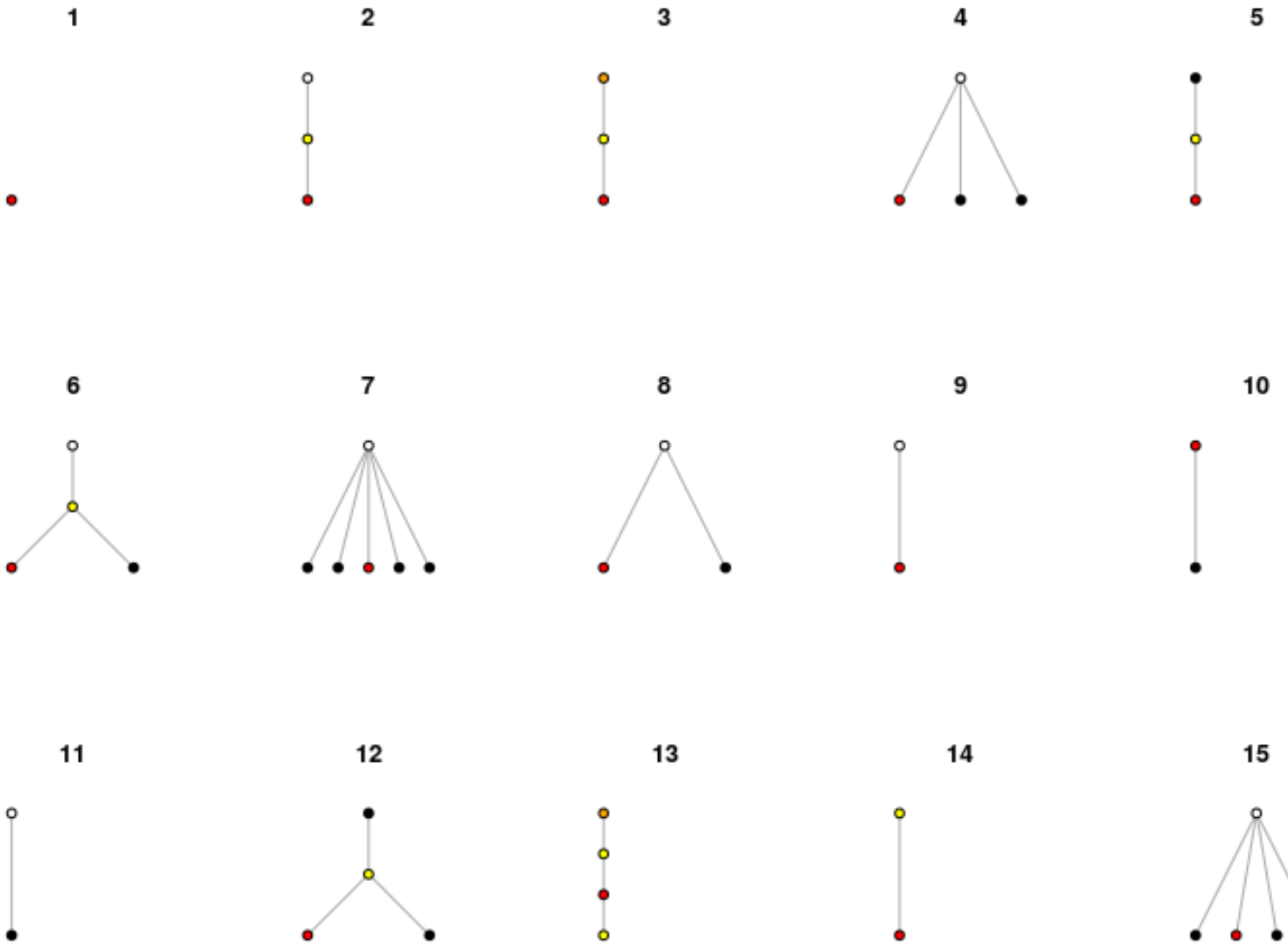
Colouring

Label special nodes:

- Red: ego.
- Yellow: parent of ego (and posts of same author)
- Orange: other posts by ego author
- White: root



Some frequent neighbourhoods



Some frequent neighbourhoods

16

17

18

19

20

21

22

23

24

25

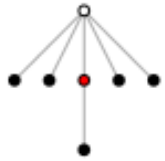
26

27

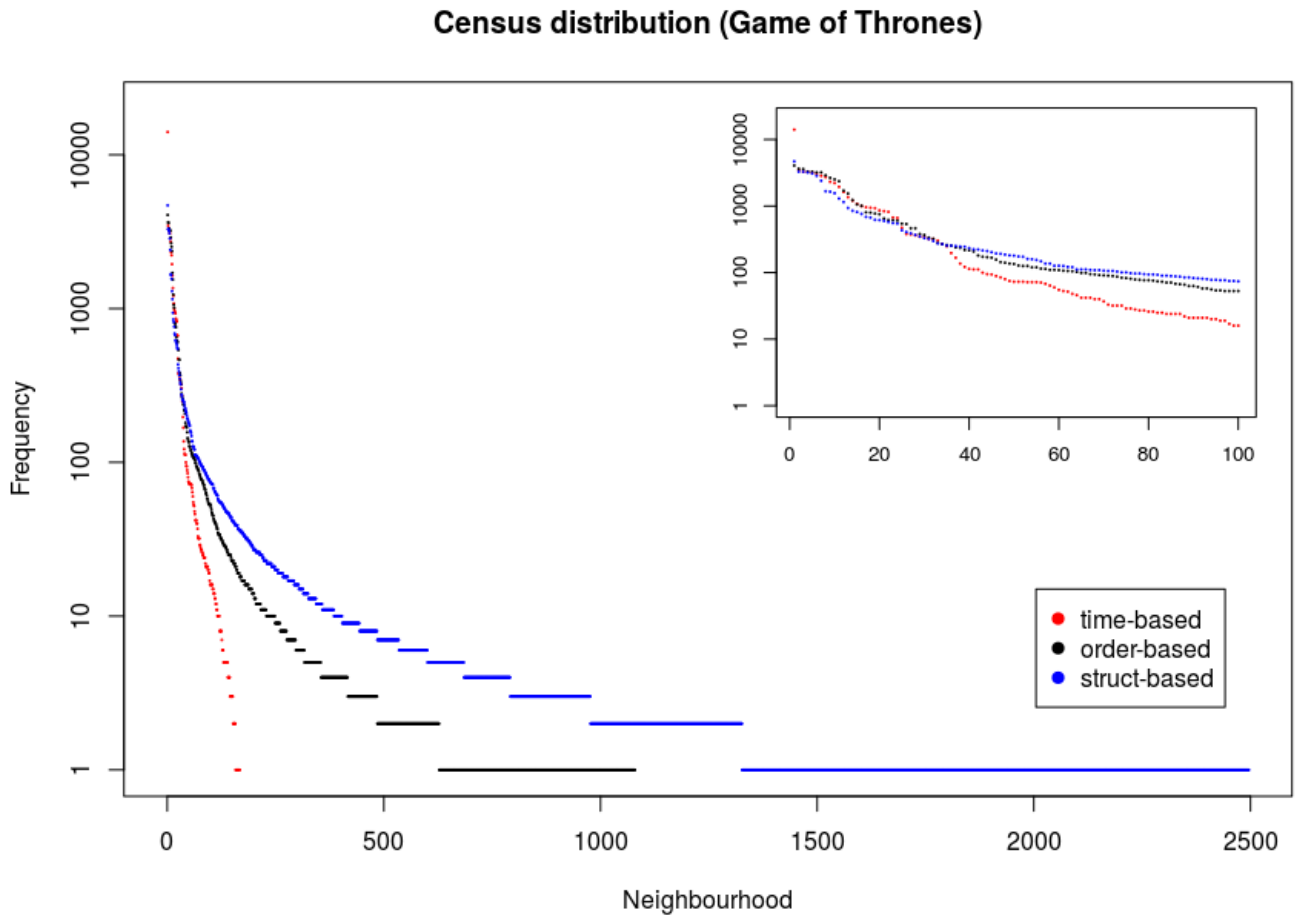
28

29

30

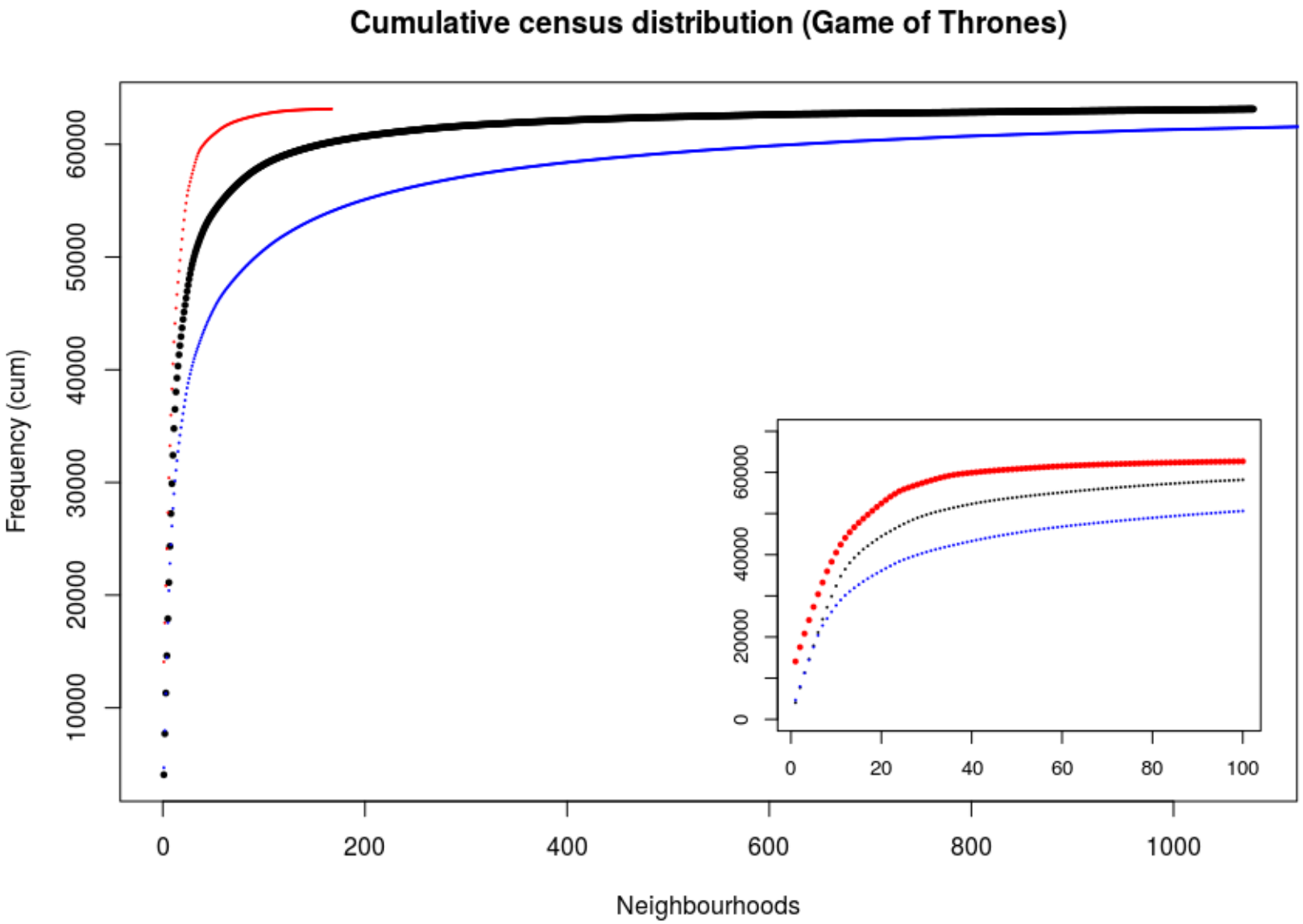


Frequency distribution



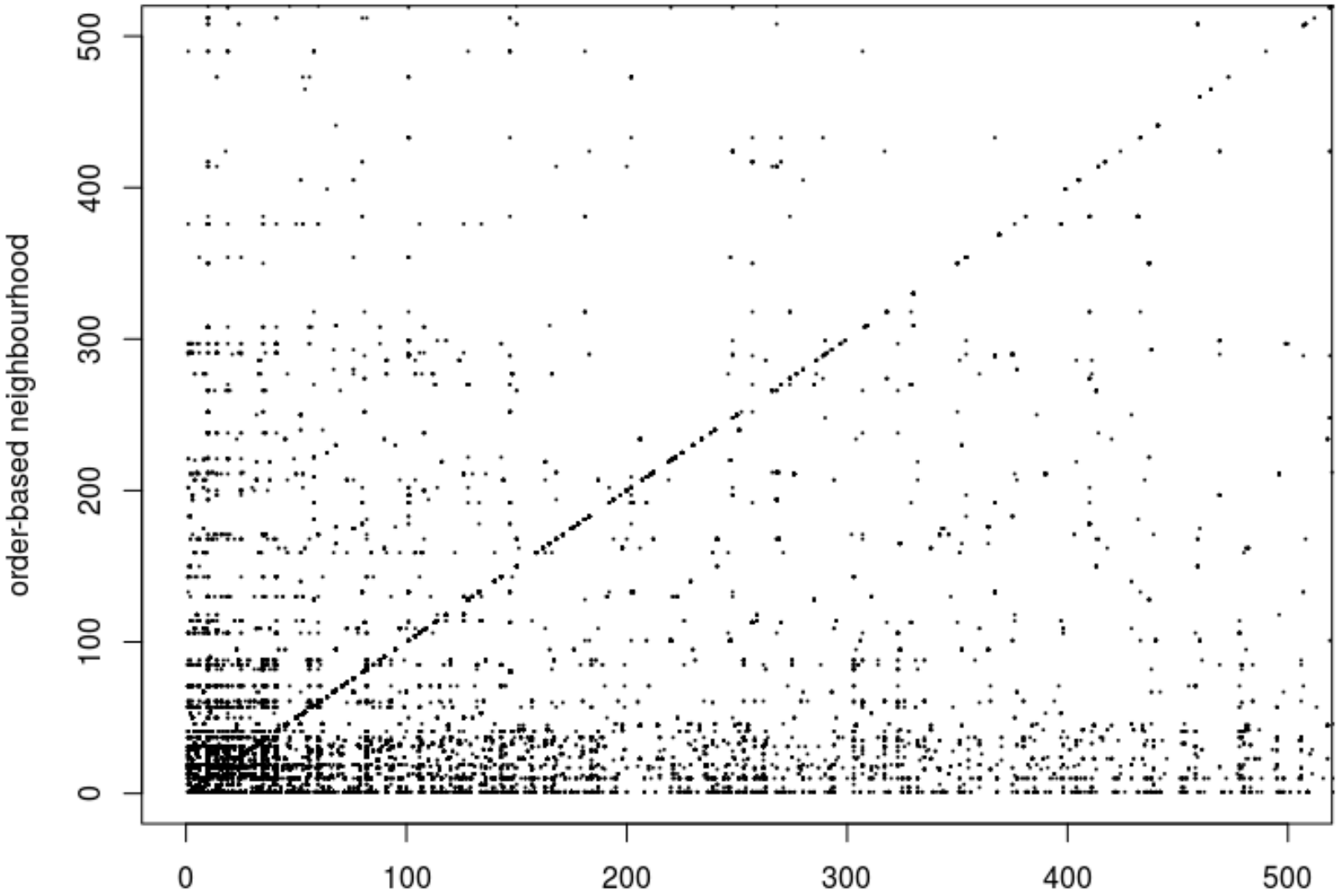
Time-based (black) reduces the space w.r.t structure-based (blue)

Frequency distribution



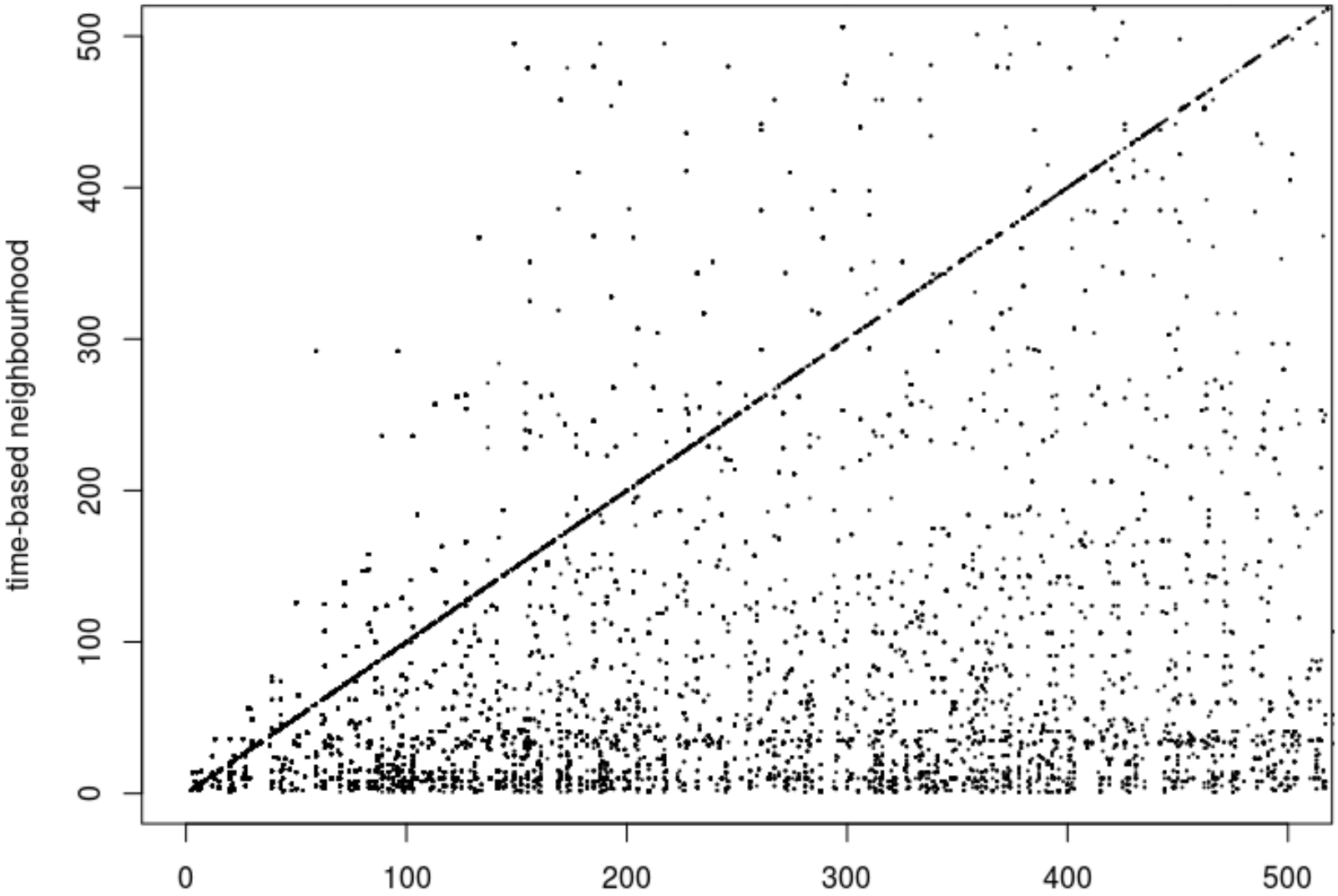
Discrepancies

Time-based vs Order-based neighbourhood



Discrepancies

Basic neighbourhood vs Time-based neighbourhood



Structure-based vs Order-based vs Time-based

Structure-based:

- too big (and too many) neighbourhoods.

Order-based:

- Dominance of monoid hides richer conversational structures.

Time-based:

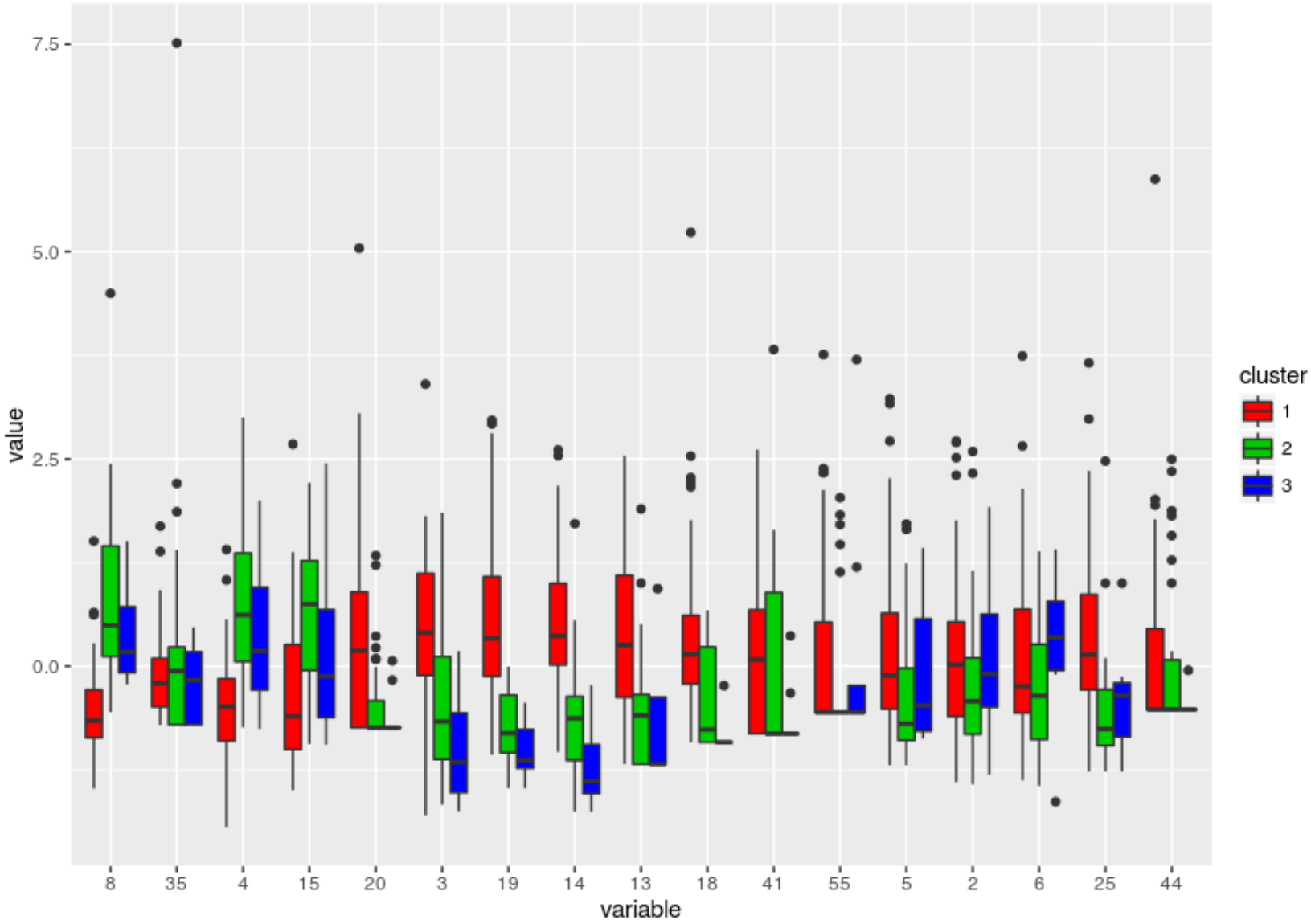
- Space more reduced than simple structure-based.
- Criteria to choose radius dynamically (r = until conversation slows down)

Methodology

- Create a user $\times$ neighborhood matrix of counts.
- Z-normalize (users characterized by their deviation from the mean)
- Cluster!

Conversation-based clustering

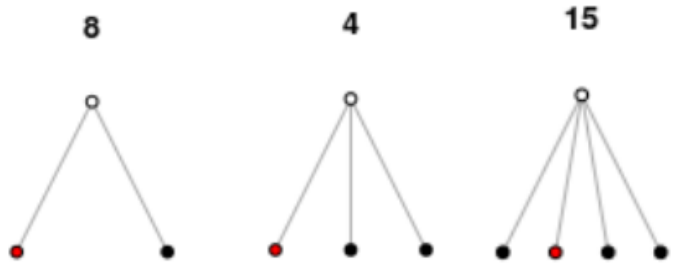
Time-based



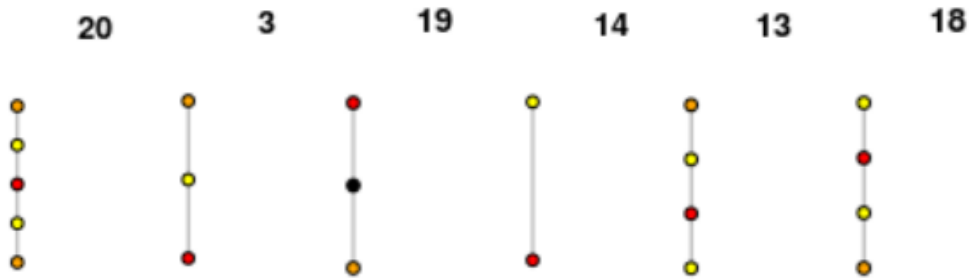
Conversation-based clustering

Interpretation

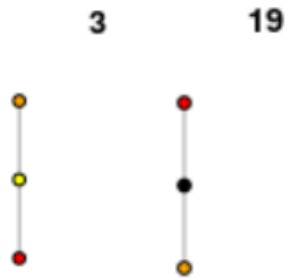
Greens:



Reds:



Blues (avoid these motifs):



Conclusions

- **Q: Can we use graph structure to characterise users?**
- A: Yes!

- **Q: By using triads**?
- A: No. They are not useful in trees.

- **Q: So, what kind of structure?**
- A: Posts neighbourhoods that are time/order sensitive.

- **Q: What about language?**
- A: It's ok, but structure is more directly linked to thread dynamics (future work)

Merci !