Measuring linguistic style alignment

Social & psychological perspectives

Bill Noble

CLASP seminar

31 October, 2018

Part I: Linguistic style alignment

- What is linguistic alignment?
- Marker-based linguistic style alignment
- Measures & sociolinguistic applications

Part II: Linguistic alignment in social networks

- Communication networks
- Network centrality measures
- Centrality and alignment
- Future work

What is linguistic alignment?

Speakers to adapt to each other at different levels:

- phonetic production (Babel 2012, Kim et al., 2011)
- lexical choice (Brennan and Clark, 1996)
- syntactic constructions (Pickering and Ferreira, 2008)

What causes this adaptation is a matter of debate:

- The need for mutual understanding (Clark, 1996)
- Low-level psychological priming (Pickering & Garrod, 2004)
- Negotiation of social distance (Gallios & Giles, 2008)

- Babel, Molly. *Evidence for Phonetic and Social Selectivity in Spontaneous Phonetic Imitation*. (2012)
- Kim, Midam, William S. Horton, and Ann R. Bradlow. *Phonetic Convergence in Spontaneous Conversations as a Function of Interlocutor Language Distance*. (2011)
- Brennan, Susan E, and Herbert H Clark. *Conceptual Pacts and Lexical Choice in Conversation*. (1996)
- Pickering, Martin J., and Victor S. Ferreira. *Structural Priming: A Critical Review*. (2008)
- Clark, Herbert H. *Using Language*. (1996)
- Pickering, Martin J., and Simon Garrod. *The Interactive-Alignment Model*. (2004)
- Gallois, Cindy, and Howard Giles. *Communication Accommodation Theory*. (2008)

Function words as style markers

- low semantic value
- usually "avoidable"
- not sensitive to topic

Category	Examples
Personal pronouns	*I, his, their*
Impersonal pronouns	*it, that, anything*
Articles	*a, an, the*
Conjunctions	*and, but, because*
Prepositions	*in, under, about*
Auxiliary verbs	*shall, be, was*
High-frequency adverbs	*very, rather, just*
Negations	*no, not, never*
Quantifiers	*much, few, lots*

Tausczik, Yla R., and James W. Pennebaker. *The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods*. (2010)

Example: Marker-based alignment

A: Corrected. Please check. Any more outstanding problems?

B: Everything is fine. Thanks a lot.

Speaker B coordinates along marker: quantifier

A: Thanks - I'll look at these over the next day or two but busy tonight.

B: OK, I'll go and do something else for the next couple of days.

Speaker B coordinates along markers: personal pronoun, preposition, article, conjunction

Metric desiderata

- accounts for speaker baselines
- accounts for message length effects
- consistent across markers
- robust to sparse data
- directional (alignment of speaker *b towards a*)
- group aggregate or individual

Style Matching (LSM) (Ireland, et al., 2011)

For a given dialogue between speakers **a** and **b** let $\mathcal{F}_{\mathbf{a}}(m)$ and $\mathcal{F}_{\mathbf{b}}(m)$ be the frequency (with respect to all tokens in the dialouge) with which **a** and **b** use marker m.

$$LSM_m(\mathbf{a}, \mathbf{b}) = \frac{|\mathcal{F}_{\mathbf{a}}(m) - \mathcal{F}_{\mathbf{b}}(m)|}{\mathcal{F}_{\mathbf{a}}(m) + \mathcal{F}_{\mathbf{b}}(m) + \epsilon}$$

and

$$LSM(\mathbf{a}, \mathbf{b}) = \sum_{m \in M} LSM_m(\mathbf{a}, \mathbf{b})$$

- score is bounded by (0,1) – higher means more "alignment"
- doesn't account for speaker baselines (sensitive to homophily)
- doesn't compare across markers
- not directional ($LSM(a, b) = LSM(b, a)$)

Speed dating (Ireland, et al., 2011)

- 40 speed dates (pairs of college students) selected for transcription
- No participant included in more than one of the selected dates
- Participants completed a "percieved similarity" survey
- Within 24 hours reported whether they would or would not be interested in a second date

Results:

- LSM predicted significantly relationship initiation ($p = 0.039$)
- daters were more than 3 times as likely to match for every standard-deviation increase in LSM ($OR = 3.05$)
- LSM remains predictive when taking into account verbosity (word count)

But is this just homophily?

Subtractive Conditional Probability (SCP)
(Danescu-Niculescu-Mizil, et al. 2012)

For speaker **b** and group **A** and reply pairs $(a_1, b_1), ..., (a_n, b_n)$, let $\mathcal{E}^m_{a_i}$ mean that utterance a_i exhibits marker m.

$$SCP^m(\mathbf{b}, \mathbf{A}) = P[\mathcal{E}^m_b \mid \mathcal{E}^m_a] - P[\mathcal{E}^m_b]$$

and

$$SCP^m(\mathbf{B}, \mathbf{A}) = \frac{\sum_{b \in B} SCP^m(\mathbf{b}, \mathbf{A})}{|\mathbf{B}|}$$

- captures directional alignment
- accounts for speaker baselines
- still doesn't compare across markers
- sensitive to utterance length

Wikipedia editor discussions
(Danescu-Niculescu-Mizil, et al. 2012)

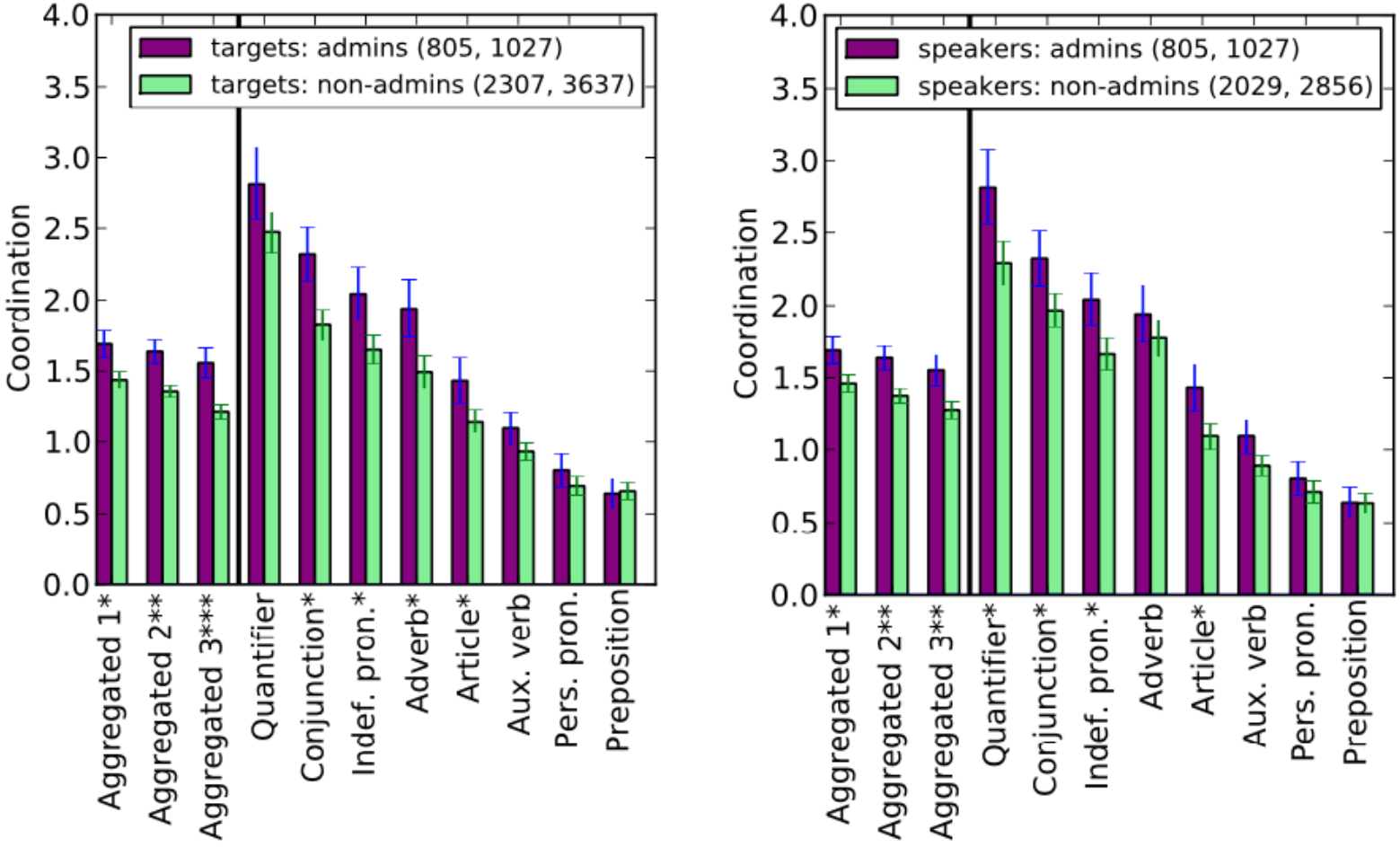


Figure: Alignment (SCP) and explicit social status

The Word-based Hierarchical Alignment Model (WHAM) (Doyle et al., 2016)

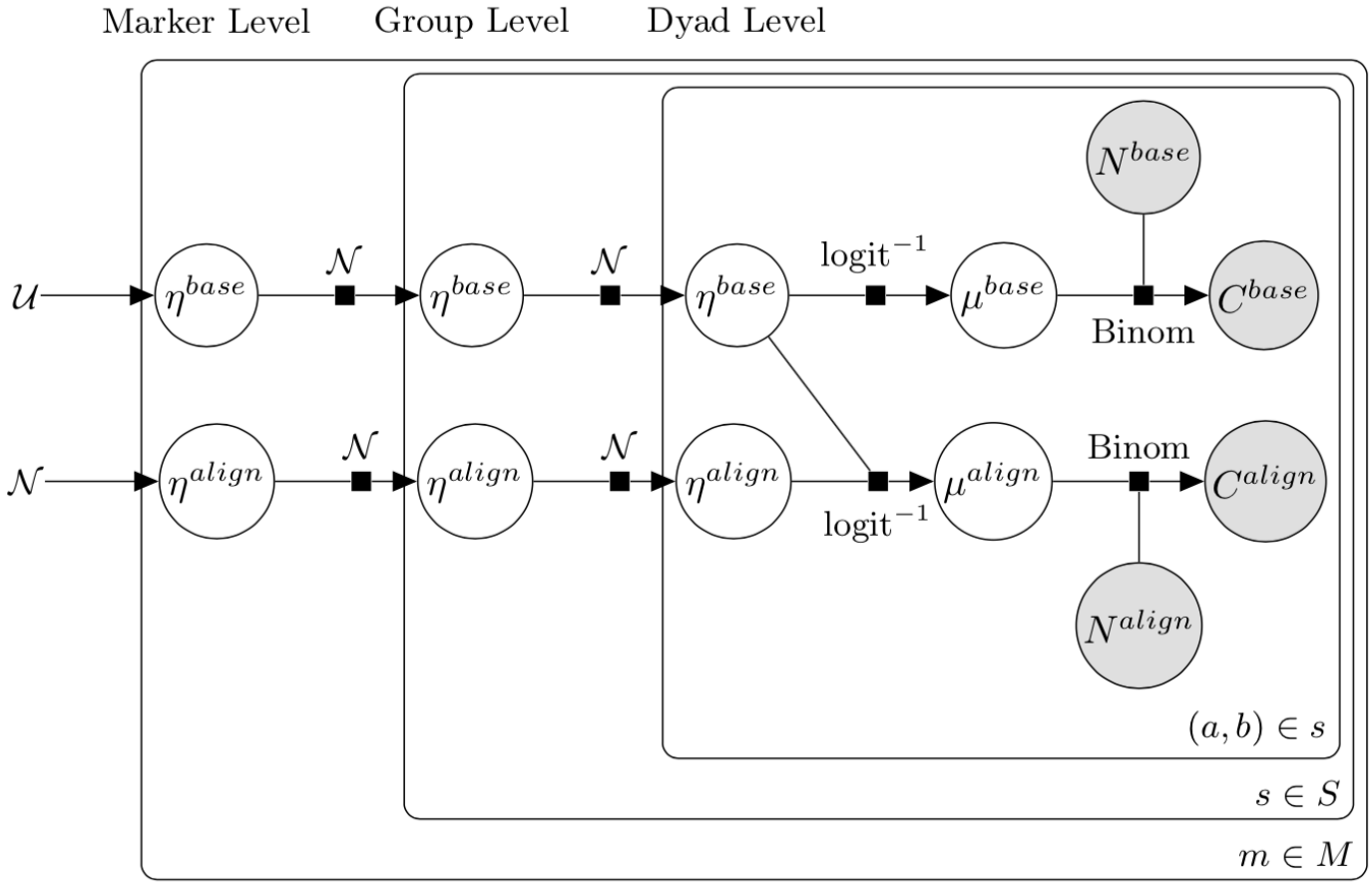
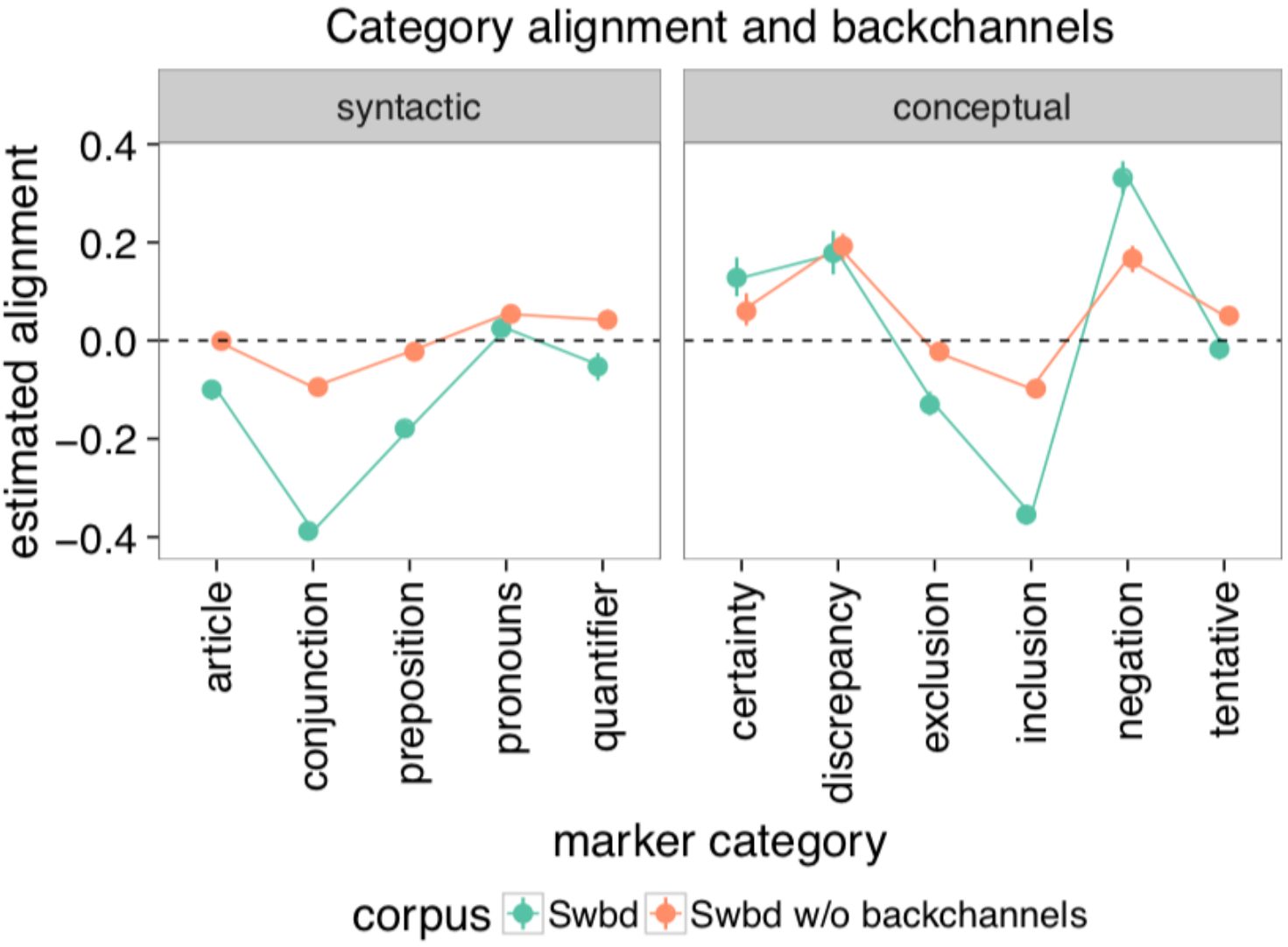


Figure: A chain of normal distributions generates a linear predictor η, which is converted into a probability μ for binomial draws of marker presence/absence

Switchboard dialogue acts

(Doyle et al., 2016)



Criticism

- linguistic alignmet is better explained by low-level features and automatic priming than social factors
- simple generalized linear model with 3 predictors: marker count (in the preceding utterance), social power (of the previous speaker), and *utterance length*
- utterance length is a low-level linguistic feature that correlates with many of the psychological causes of alignment

Xu, Yang, Jeremy Cole, and David Reitter. *Not That Much Power: Linguistic Alignment Is Influenced More by Low-Level Linguistic Features Rather than Social Power*. (2018)

Part I: Linguistic style alignment

- What is linguistic alignment?
- Marker-based linguistic style alignment
- Measures & sociolinguistic applications

Part II: Linguistic alignment in social networks

- Communication networks
- Network centrality measures
- Centrality and alignment
- Future work

Communication networks (social networks for linguists)

We can model a speech community as a social network:

- Nodes are speakers
- Edges represent (some measure of) communication between them

For the Wikipedia Talkpages Corpus (from *Echoes of Power*):

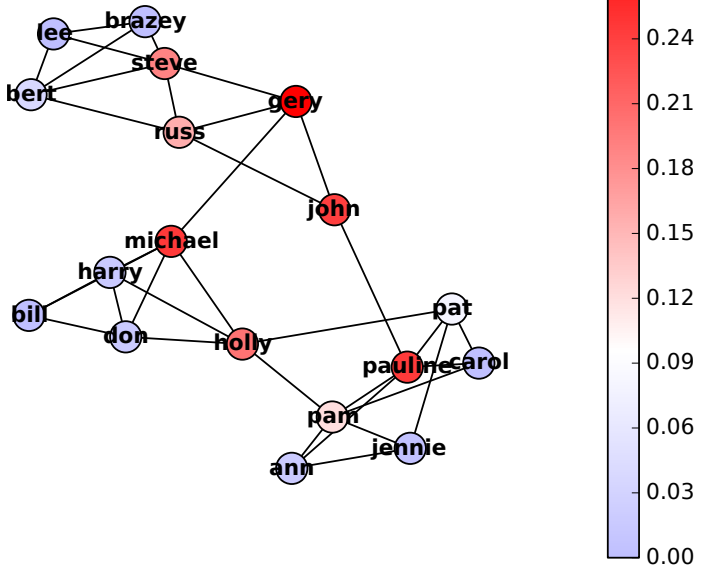
- Nodes are Wikipedia editors.
- Edges are weighted according to the number of direct talkpage replies between editors.
- Edges are undirected (although, this is a choice).
- Total of 25826 nodes, 85731 edges.
- Average degree (number of neighbors) $= 6.64$

Betweenness Centrality

How important are you to community connectivity?

$$\mathrm{BC}(n^*) = \sum_{n \neq m \in N} \frac{|\{\sigma \in \mathrm{Path}(m,n) \mid n^* \in \sigma\}|}{|\mathrm{Path}(m,n)|}$$

where $\mathrm{Path}(m,n)$ is the set of shortest paths between m and n.

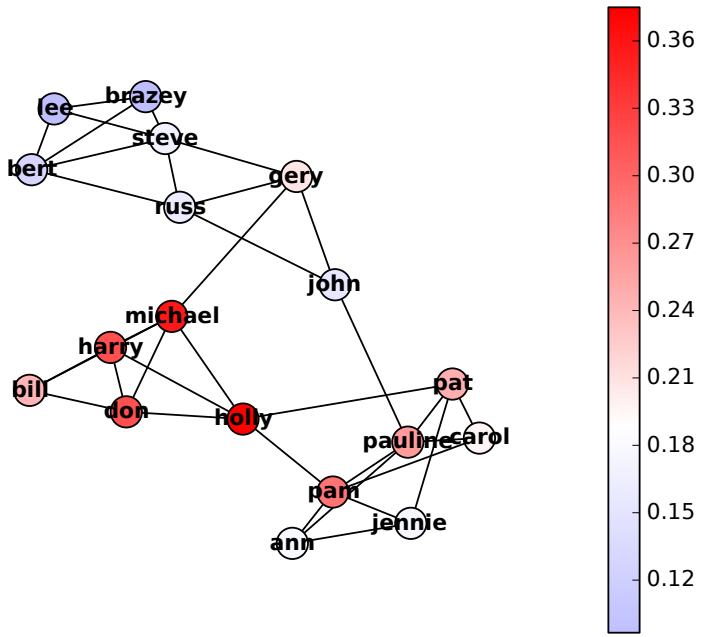


Eigenvector Centrality

How important are your neighbors?

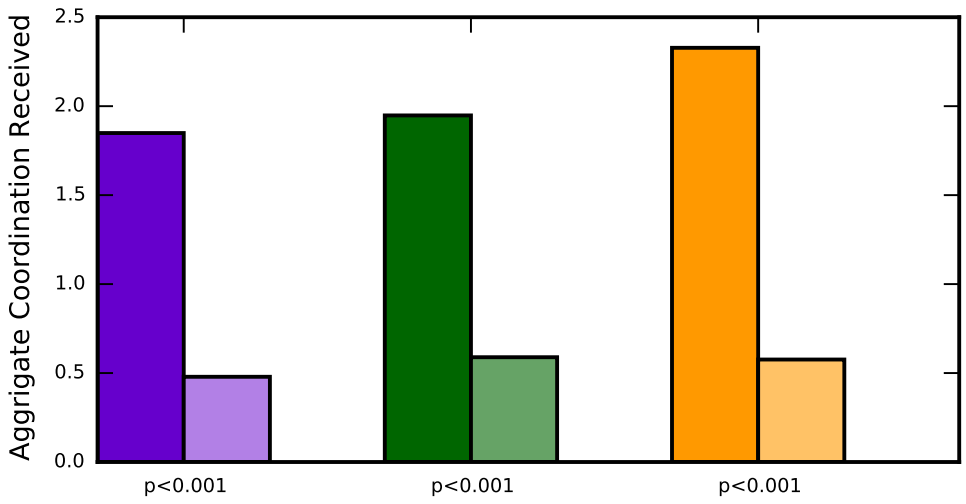
$$\mathrm{EC}(n^*) = \frac{1}{\lambda} \sum_{n \in M(n^*)} \mathrm{EC}(n)$$

where $M(n)$ is the neighborhood of n and λ is the largest *eigenvalue*



Centrality and style alignment

Highly central editors receive more coordination.

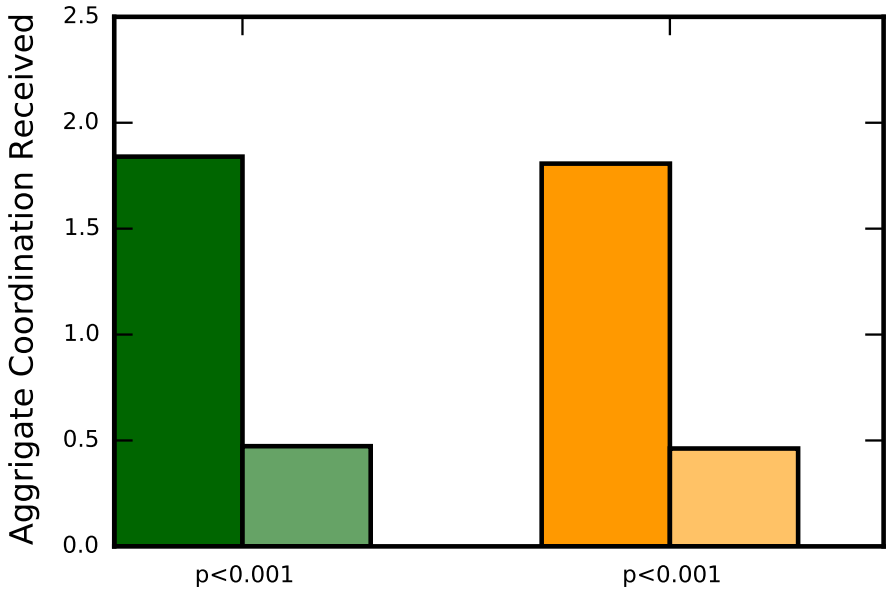


- Admins
- Non-Admins
- High Eigenvector
- Low Eigenvector
- High Betweenness
- Low Betweenness

(Noble & Fernández, 2015)

Centrality and alignment

Low-centrality users receive more coordination if they are admins.

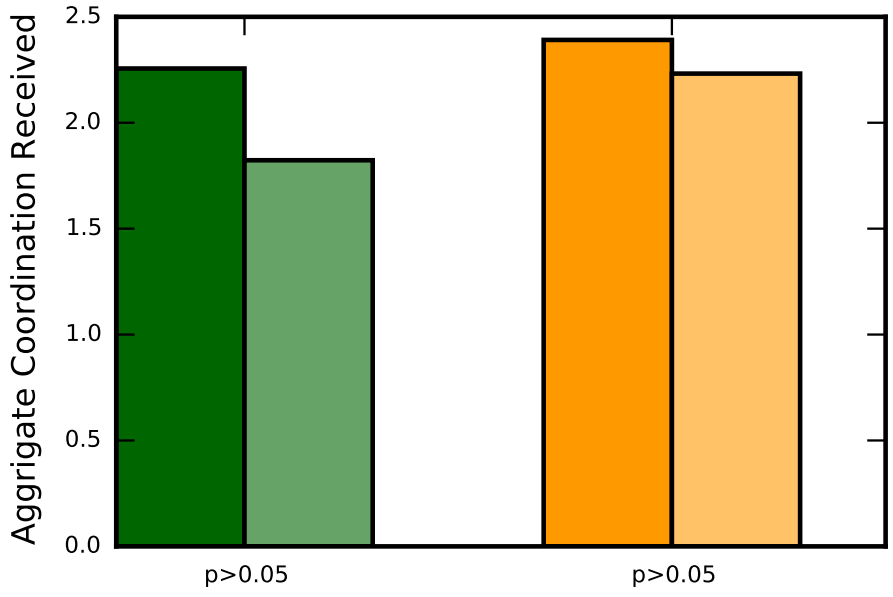


- Low Eigenvector Admins
- Low Eigenvector Non-Admins
- Low Betweenness Admins
- Low Betweenness Non-Admins

(Noble & Fernández, 2015)

Centrality and alignment

Adminship is less important for high-centrality users.



- High Eigenvector Admins
- High Eigenvector Non-Admins
- High Betweenness Admins
- High Betweenness Non-Admins

(Noble & Fernández, 2015)

Future (current) work

- More domains (citizen science forums, for example)
- Use WHAM (or improve SCP)
- Investigate other social network features (Louvian sub-communities, for example)
- Probe the sources of linguistic style alignment (does stylistic typicality play a role?)