UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Welcome
Bienvenue
Benvenuti
Willkommen

Computer-assisted text analysis to measure legitimacy in social media

Keywords in context (KWIC)

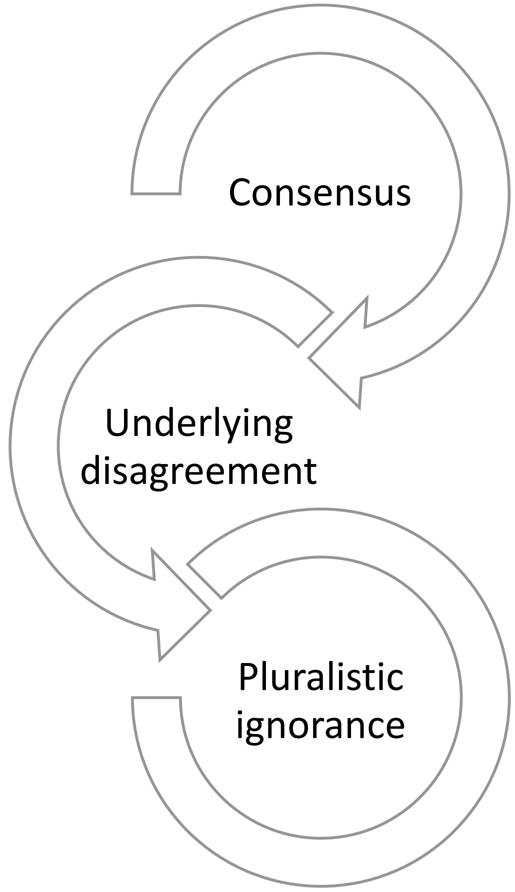
Paper Development Workshop & Expert Panel on Legitimacy Theory, Lausanne, 8-9 May 2023

Prof. Dr. Laura Illia, Chair of Communication, Business and Social Responsibility, Department of Communication and Media Research (DCM) Faculty of Management, Economic and Social science, Laura.illia@unifr.ch

Advancing the Multilevel Theory of Legitimacy: Relevance of social media

Ample opportunities to advance ...

"We see ample opportunities to advance our understanding of legitimacy by drawing on the context of social media (...) The rise of social media as a heterogeneous and coproduced environment changes how social judgments about organizations are produced and disseminated (...) may help flag abrupt institutional change (...) can also provide visible evidence of cascades that reinforce legitimacy, destroy it, or create something new." (Haack et al., 2021, p. 24)



Context of Social Media: upward loop

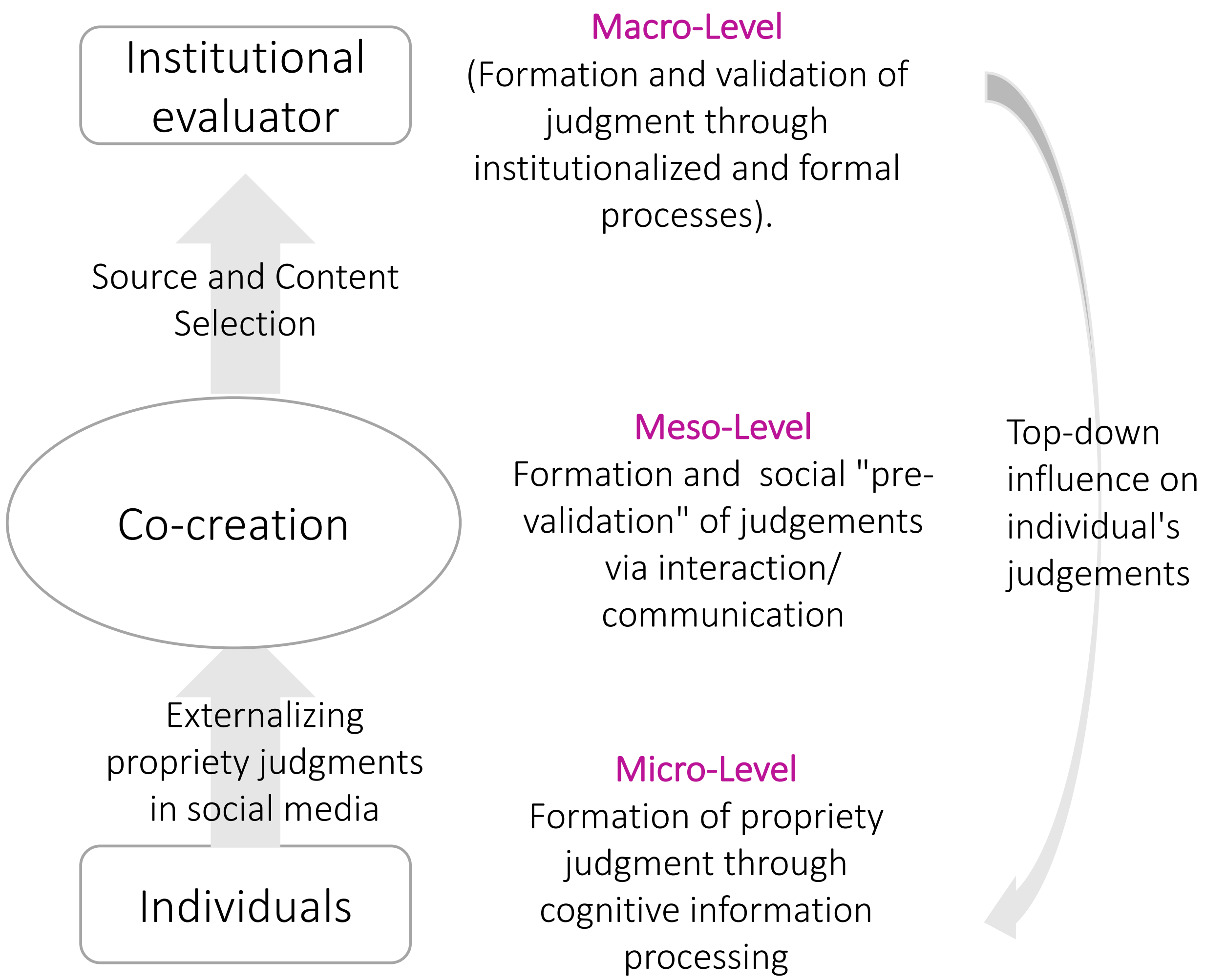
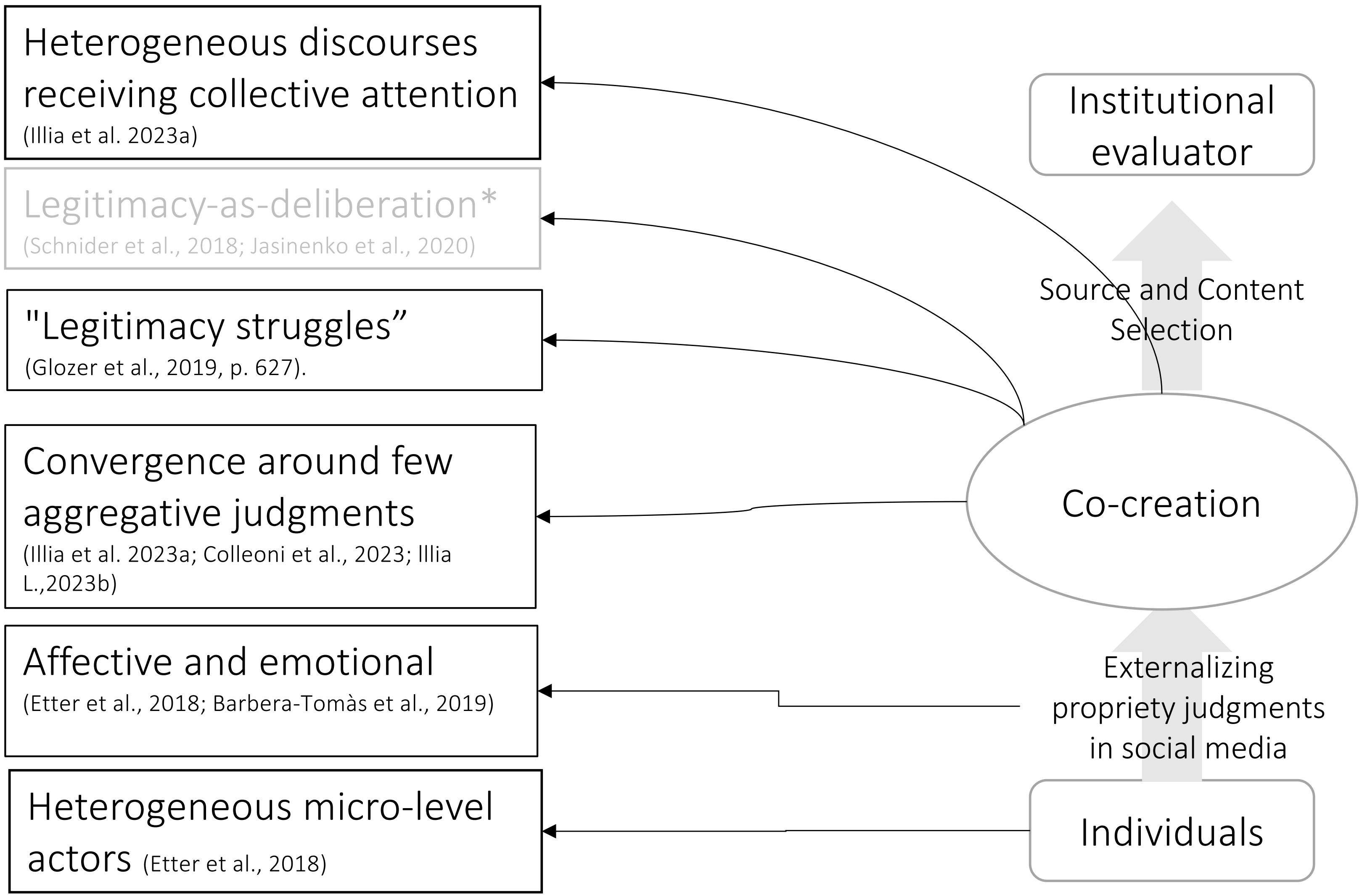


Figure source :Adapted from Illia, Etter, Meggiorin, Colleoni ,2022

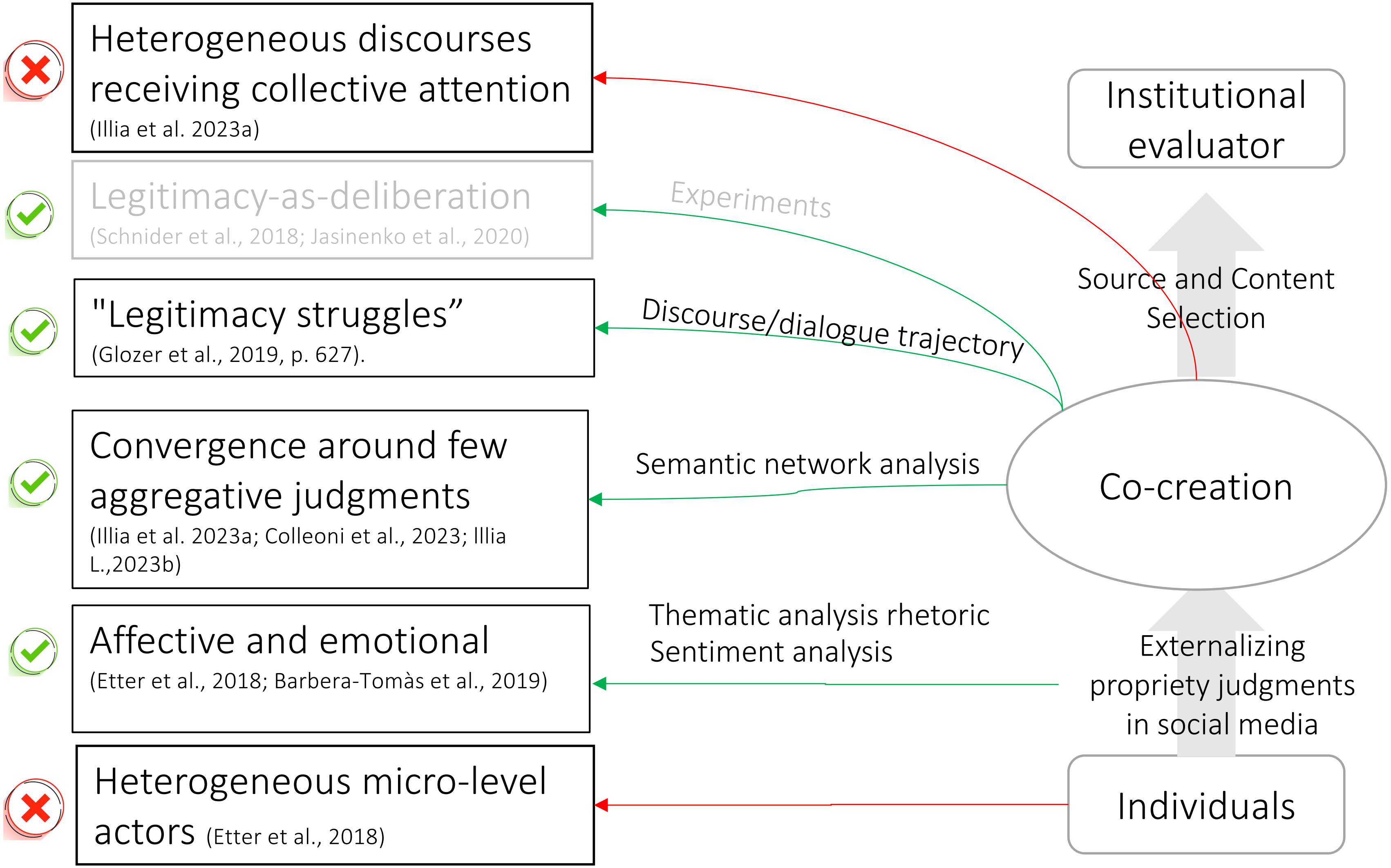
What to analyze in this upward loop?

Extant studies have analysed...



*Not yet used to analyze social media dynamics in legitimacy studies but has excellent potential as deliberative experiments may create interactions-communications in communities and allow synchronic deliberation.

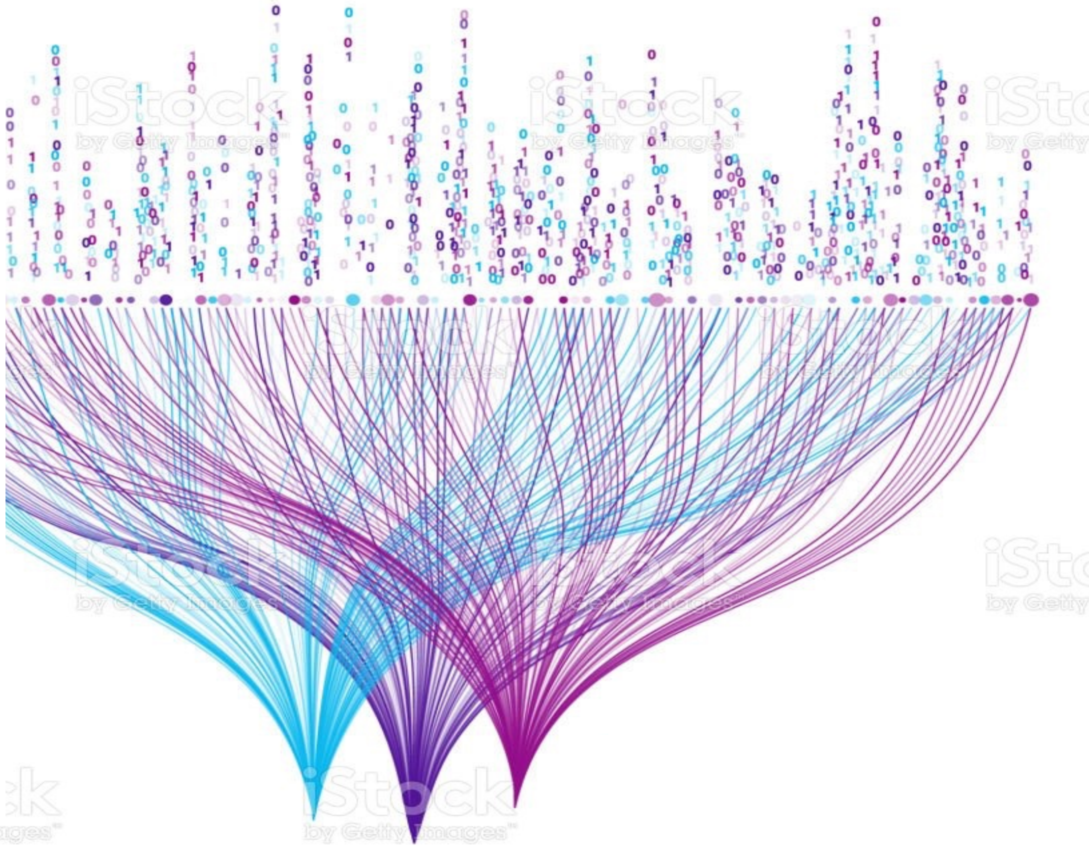
Which methods at hand...?



Why at this stage?...Because data look like.....

Who is (de-) legitimizing? → Which discourses receive collective attention?

Main issue: when micro-actors interact as a group, the group is informal –most if not all times temporary. Hence, one can identify who they are on the base of what they say. Yet, identifying "who is who" on the basis of what they say in a tweet would be problematic, tautological.....



Computer-assisted quantitative text analysis: its benefits and limitations

Computer-assisted quantitative text analysis

"refers to a range of current techniques from quantitative social science and content analysis to data mining and text classification (of) data in the form of natural language texts of social scientific interest" (Bauer et al., 2014)

Approach	Method	What it does
Positioning	Keyword-in-context **(KWIC)**	Produces **meaning of word is provided by context where it is inserted**
Inference	Keyword-out-of-context **(KWOC)** – **dictionary based**	Assigns **meaning by inferring that a word has a meaning**

Computer-assisted quantitative text analysis

"refers to a range of current techniques from quantitative social science and content analysis to **data mining** and text classification (of) data in the form of natural language texts of social scientific interest" (Bauer et al., 2014)

Approach	Method	What it does
Positioning	Keyword-in-context **(KWIC)**	Produces **meaning of word is provided by context where it is inserted**
Inference	Keyword-out-of-	Assigns meaning by inferring that … ing

The general approach of this method is: words co-occurring in relative proximity (i.e., distance) in the same context (i.e., text-corpus) are interpreted as relating to a common theme or concept in the discourse studied.

Key Word in Context (KWIC)

Advantages

Allows to identify hidden patterns

- ✓ "that are generally impossible to specify a priori" (Alemany & Vayre, 2015:11)
- ✓ that are impossible to be grasped with 'human coding' due to a high volume of data (Baumer, Minmo, Guha et al., 2017).
- ✓ that can enable an excellent qualitative analysis (Baumer, Minmo, Guha et al., 2017)

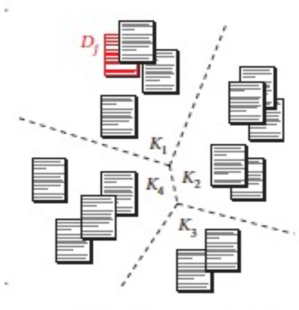
Disadvantages

Does not allow us...

- X to provide theoretical advancement unless one also conducts inferential tests (quantitative) or theory building (qualitative) (Aranda et al., 2021; Hannigan et al., 2019) .
- X Interpretation of machine learning indicators is tidy (Hannigan et al., 2019).

Examples KWIC used so far for legitimacy

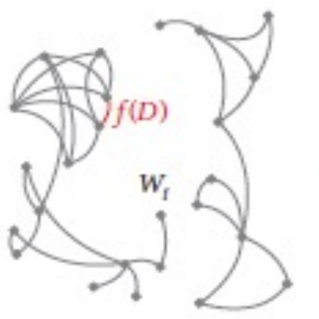
Sentiment
(machine learning)



Categorical
distances

Judgments'
valence
character

Semantic
Network



Topological
distance

Judgment'
attributes
Heterogeneity
vs Convergence

Visualizations adapted from Evans and Aceves, 2016: :21-50

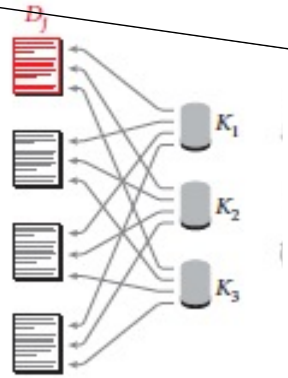
Examples KWIC NOT used so far for legitimacy

Analyse user's descriptions

Analyse tweets

Extant studies still have not proposed this. STM and VSM are ideal, respectively for:

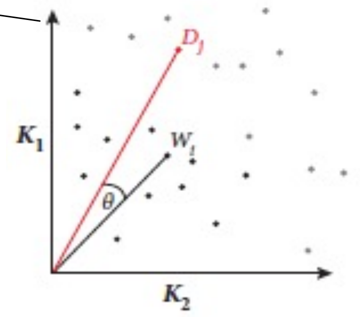
Structural Topic Modeling (STM)



Combinational distance

Attributes of heterogeneous discourses receiving attention

Vector Space Modeling (VSM)



Geometrical distance

Heterogeneous micro-actors and their identifying attributes

Visualizations adapted from Evans and Aceves, 2016: :21-50

STM vs VSM at a glance

	STM (Roberts et a.., 2016;2019)	VSM (Mikolov et al., 2013)
Distance	Combinational	Geometrical
Statistical model	Bayesian probabilistic model (LDA)	Word embedding model (NLP)
Focus is	Semantic description of attributes of a salient discourse	Semantic distance (discourse attributes similarities vs differences)
Used to trace..	Collective attention over certain discourses attributes: probabilistic model calculates most prevailing topics (i.e., discourses) in a dataset based on words (FREX) most significantly related to a topic.	Social/cultural dimension of discourses attribute: NLP models used for VSN grasp important semantics typical of a specific culture/social system
Disadvantage	Corpus text sparsely distributed, as one document (e.g. tweet) belongs to more topics (threshold for exclusiveness difficult to be decided)	Pre-training dataset has to be hudge (NLP). Otherwhise emerging clusters are nonesense.
Advantage	Meta-variables to do inferences, and allow causal relationships and test hypothesis or build theory	Corpus-text exclusively embedded as one document (e.g.Tweet) belong to one cluster only.

Attributes of heterogeneous discourses receiving attention

Heterogeneous micro-actors and their identifying attributes

Example of application:

Financed by

Becas Leonardo
a Investigadores y
Creadores Culturales
Fundación BBVA

Laura Illia (PI),
U Fribourg
Switzerland

Marco Caserta ,
IE University,
Spain

Nuccio Ludovico,
U Groningen,
Netherlands

Uber: Underlying disagreement voiced...

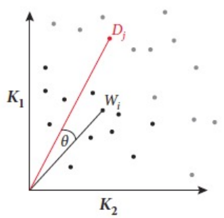
Methods: Data collection and analysis

KWIC

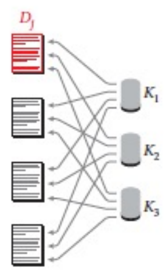
118,202 Users' descriptions

113,971 Tweets*

Vector space



Topic model



Manual coding (qualitative)

Human interpretation of resulting robust (hidden) patterns

Regression analysis

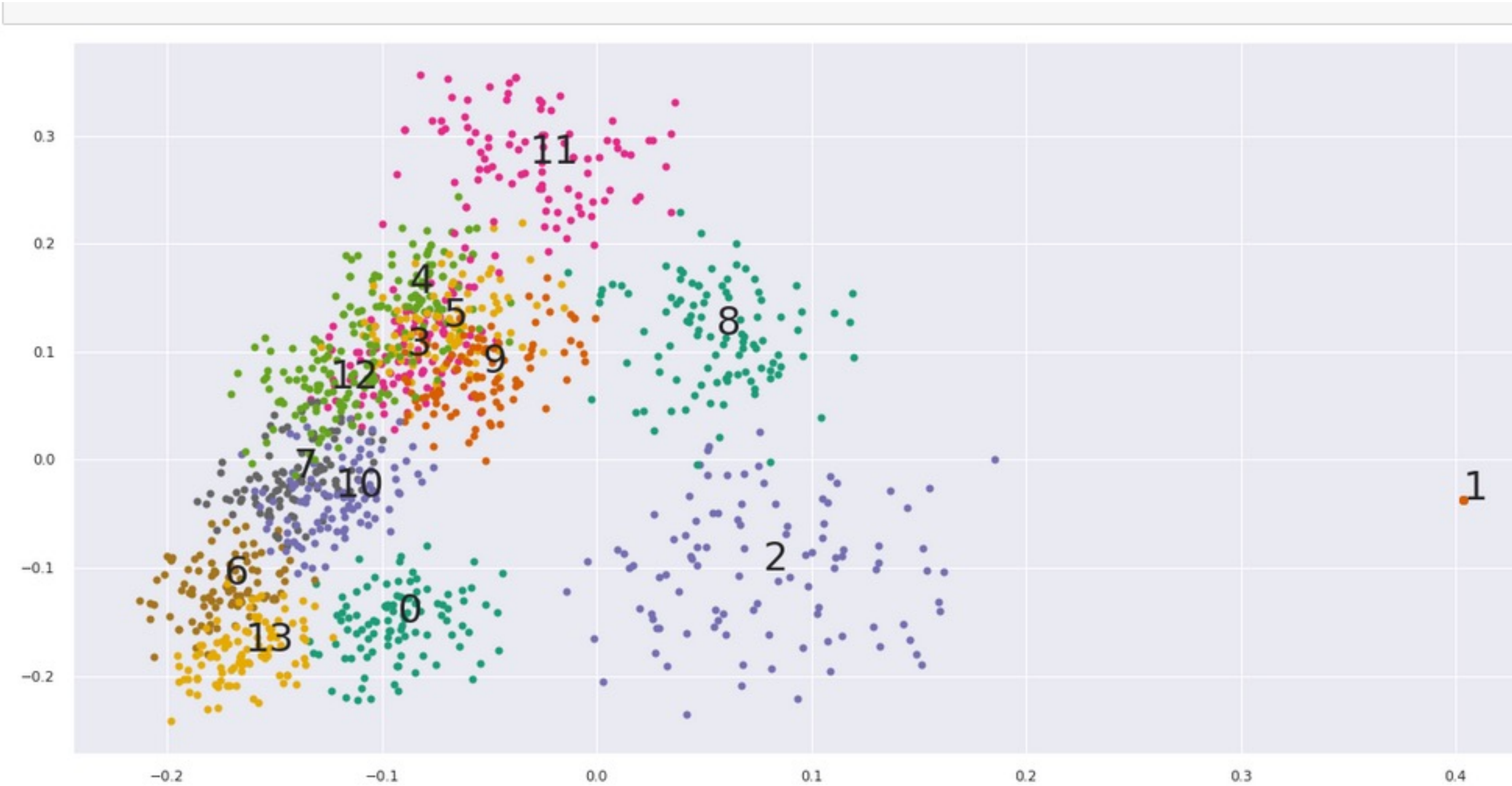
to analyze which VSM clusters significantly influence the rise of which STM topic when (time-period)

$$\hat{y} = \sum_{i=0}^{13} \beta_i \, c_i + \sum_{t=2}^{5} p_t + \sum_{t=1}^{5} \sum_{i=0}^{13} \beta_{it} c_i p_t$$

where c_i indicates the cluster number, p_t indicates the period, and c_ip_t models interaction cluster-period.

* initially 149,366 (before clearing techniques: Lemmatization (Spacy); Stopwords (1100 terms); Threshold f_{term} >= 20; wordLengths >= 4; Cleaned up punctuation)

Step 1: 118.292 users' descriptions clustered in 14 clusters (VSM)



The approach proposed is inspired by forensic linguistics: I will tell you who you are based on how you describe yourself

Qualitative analysis of user's descriptions

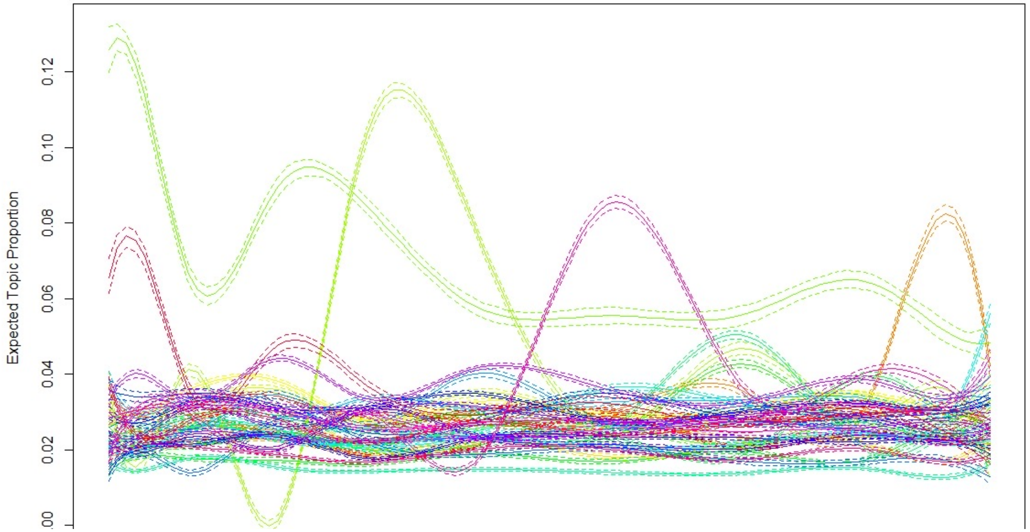
13	**Disillusioned**	Self-description is very pessimistic in terms of trust in others and reality. Average n° of followers	how should we like it were stars to burn with a passion for us we could not return if equal affection cannot be let the more loving one be me auden (user 11411332) do not draw a conclusion from me by reading my bio you are sad to love is to become vulnerable to that person that you do not know if it will end up destroying you (user 33080849)
3	**Creative and activist**	Self-description denoting profession related with creativity and art that are engaged politically. Very high n° followers	creative leader codean at adhouse advertising school illustrator podcaster new york red bulls fan father to two nincompoops black lives matter » (n° user, 489396705) best known for being mds humanitarian activist photographer goofball actor lover of cupcakes starbucks and razorback football (user 20561849)
4	**Tech professional**	Self-description denoting profession related with IT and entrepreneurship in IT. Below average n° followers	student of cities businessman consultant technologist innovator passion for gis global iop ambassador founder innorte urban expert advisor medell (user 947383580) senior software engineerteam lead by profession photographer designer twitter entrepreneur achiever (user 76000475)

Figure 3: VSM clusters

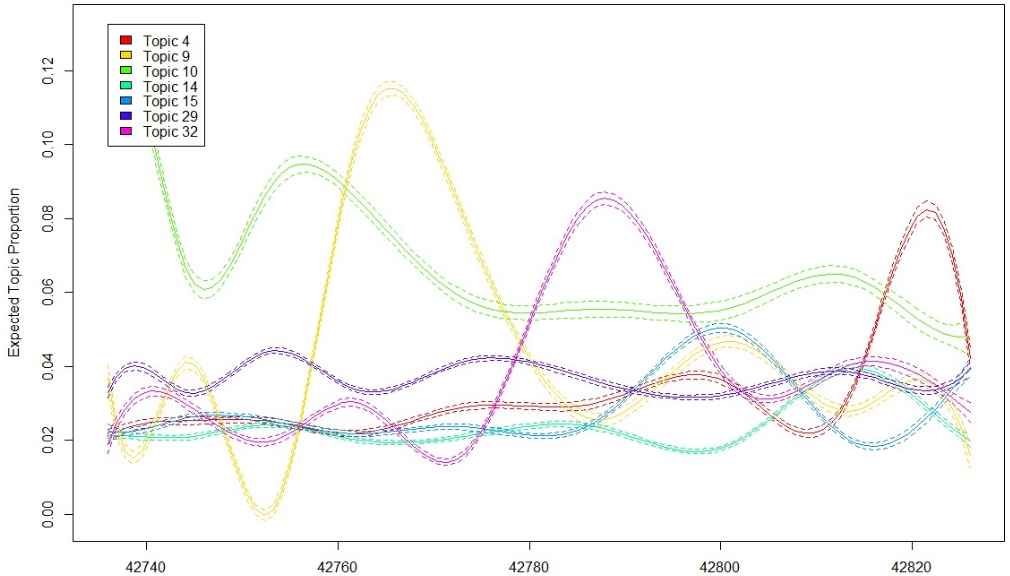
VSM cluster	Patterns in self-descriptions	VSM - Actor label
VSM cluster n° 0: 7138 users, 8445 tweets, avg n° followers 4549	Self-description showing off own diversity and uniqueness without shame; Low number of followers	The diverse
VSM cluster n° 2: 961 users, 1024 tweets, avg n° followers 21160	Self-description with expressing freedom to exert own sexuality and express own creativity and love for art. Video markers, passion for cultural industries; Low number of followers	The free creatives
VSM cluster n° 8: 3220 users, 3847 tweets, avg n° followers 2834	Self-description on the basis of slang expressions and language that refers to hashtags circulating in Twitter or mentions. Very diverse user type as they show	The hashtaggers
VSM cluster n° 11: 1820 users, 2193 tweets, avg n° followers 10116	Self-description which cites titles of tv series, song titles and use acronyms to refer to other artistic productions. Low number of followers	The acronymers
VSM cluster n° 7: 10164 users, 14834 tweets, avg n° followers 53878	Self-description based on profession in news media (news publishing, news content production). Average number of followers	News specialists
VSM cluster n° 9: 9010 users, 11367 tweets, avg n° followers 13740	Self-description denoting expertise and profession in public relations, public affairs or web communication. Below average number of followers	PR specialists
VSM cluster n° 10: 16727 users, 23777 tweets, avg n° followers 23204	Self-description denoting passion or profession in editorial or entertainment industry; Above average number of followers	Entertainment specialists
VSM cluster n° 12: 120907 users, 15589 tweets, avg n° followers 38423	Self-description denoting profession related to content production in marketing, advertising or journalism. Largely above average number of followers	Content specialists
VSM cluster n° 4: 4537 users, 5831 tweets, avg n° followers 9208	Self-description denoting profession related to IT and entrepreneurship in IT. Below average number of followers	Tech professionals
VSM cluster n° 6: 10249 users, 20792 tweets, avg n° followers 31774	Self-description very prosaic, with quotations or "maxims" that talk about life and meaning of life. Largely above average number of followers	Philosophers
VSM cluster n° 13: 9120 users, 11556 tweets, avg n° followers 10274	Self-description very pessimistic about trust in others and reality. Average number of followers	Disillusioned
VSM cluster n° 3: 6319 users, 7529 tweets, avg n° followers 3587	Self-description denoting profession related to creativity and art, engaged politically. Very high number of followers	Creative activists
VSM cluster n° 5: 4559 users, 5537 tweets, avg n° followers 9426	Self-description denoting profession related to IT and entrepreneurship in IT. Below average number of followers	Ideological hackers
VSM cluster n° 1: 23471 users, 29571 tweets, avg n° followers 2012	No self-description	Anonymous

Step 2: STM of 113,971 tweets

All 35 discourses -topics with collective attention. Here their fluctuation



Zooming into 6 topics that fluctuate the most during the boycott expressing de-legitimizing evaluations about UBER



6 discourses : qualitative analysis (e.g.)

Highest Prob	taxi, service, play, catch, bitch, hand, cabify, seat, transportation, launch, industry, competition, comment, fart, bike
FREX	service, play, catch, bitch, hand, cabify, seat, transportation, industry, competition, comment, fart, bike, american, regulate
Lift	alliance, labor, rescue, simplified, american, association, bike, bitch, blessing, boat, cheat, citizen, comment, competition, current
Score	taxi, service, play, catch, bitch, hand, cabify, seat, industry, transportation, launch, competition, fart, comment, bike

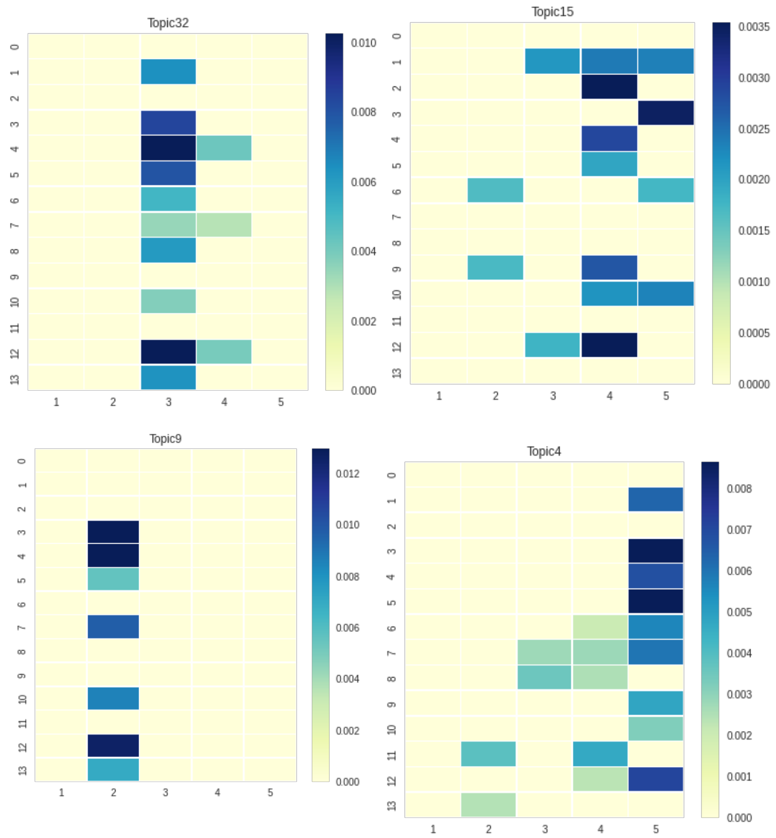
uber use simplified business model disrupt taxi industry (tweet n° 9768)
citizen request taxi driver improve provision service way taxi service sustainable (tweet n° 128936)

Highest Prob	company, tech, rate, post, harassment, sexual, sound, culture, agree, engineer, sexism, lead, past, female, billion
FREX	company, tech, rate, post, harassment, sexual, sound, culture, agree, sexism, lead, past, billion, singe, investor
Lift	annoy, curse, snap, technews, allege, artist, billion, blast, blog, candidate, careem, content, crown, culture, demonstrate
Score	company, harassment, sexual, tech, rate, post, sexism, culture, sound, engineer, investigation, resign, agree, lead, female

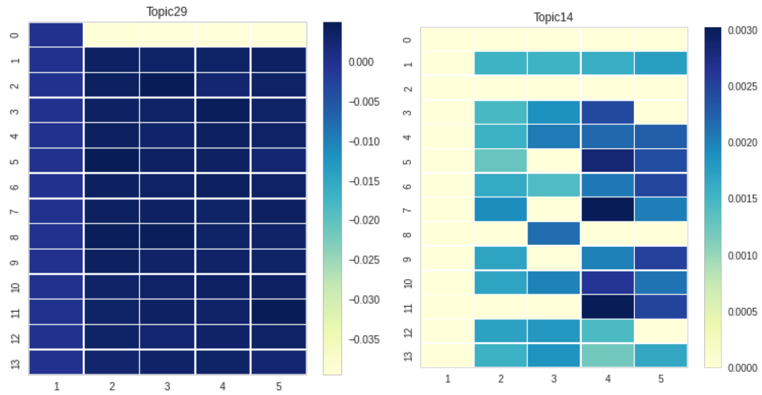
blog post write female engineer uber/describe rampant sexual harassment discrimination (tweet n° 138937)
news tech uber tech company mishandle sexual harassment claim (tweet n° 139300)

Step 3: $\hat{y} = \sum_{i=0}^{13} \beta_i c_i + \sum_{t=2}^{5} p_t + \sum_{t=1}^{5} \sum_{i=0}^{13} \beta_{it} c_i p_t$

4 discourses receiving a temporary pick of attention

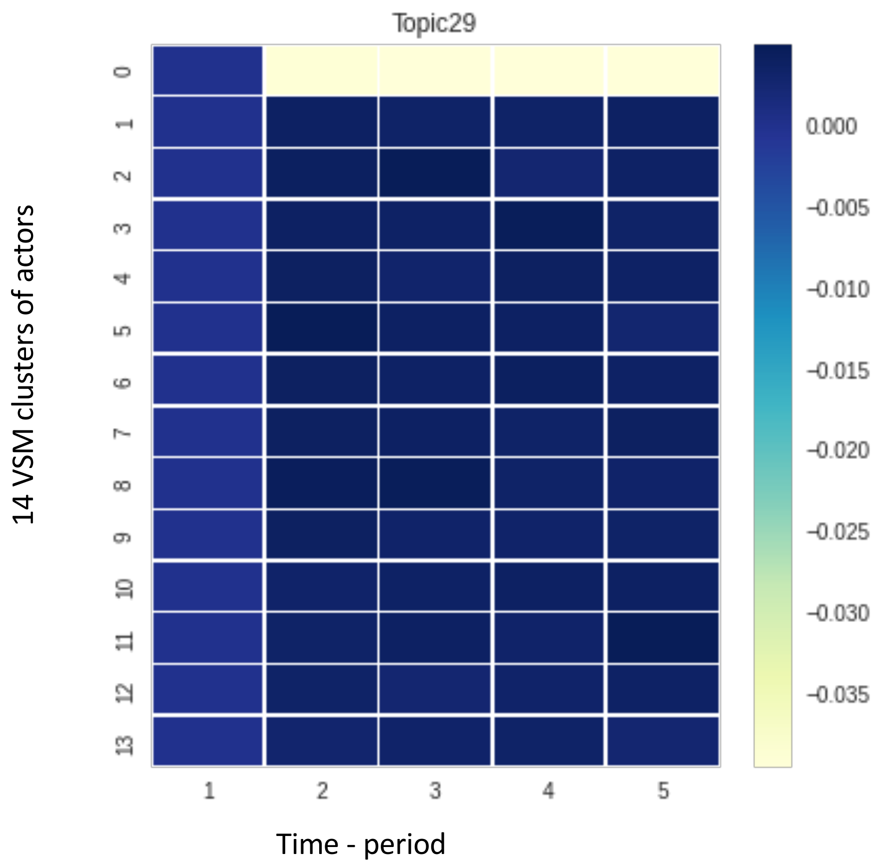


2 discourses receiving longer pick of attention

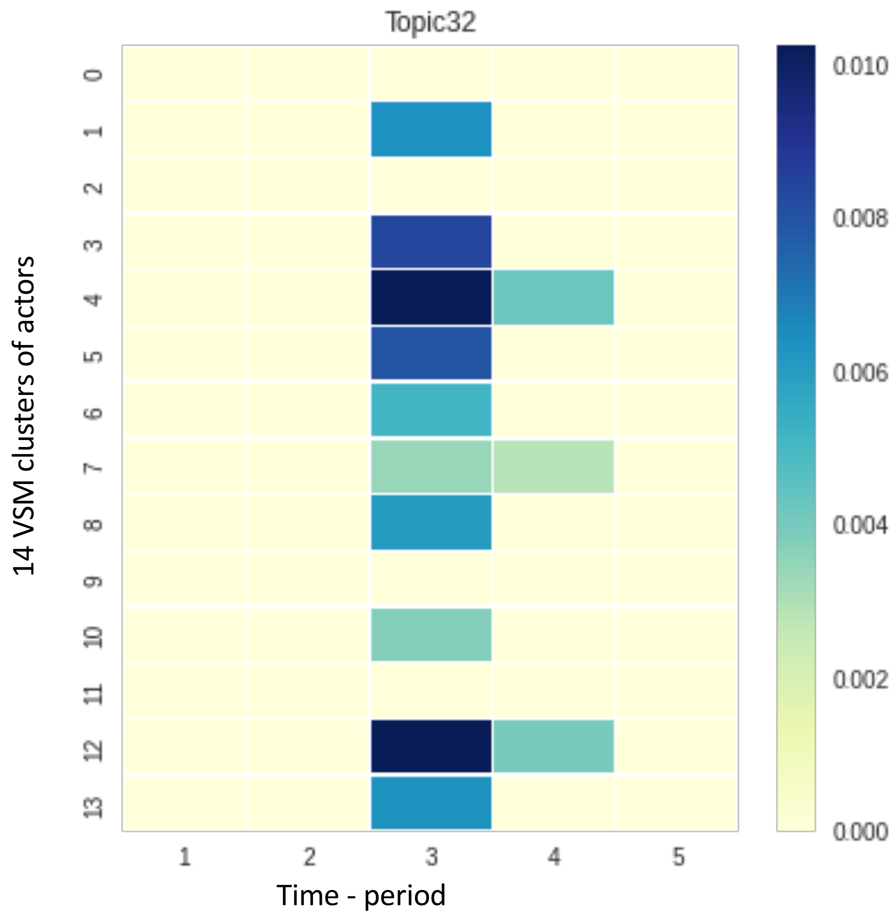


Example Discourse 29 and 32

- 29: sharing economy_ lack regulation
- 32: Sexist culture tech industry-Uber



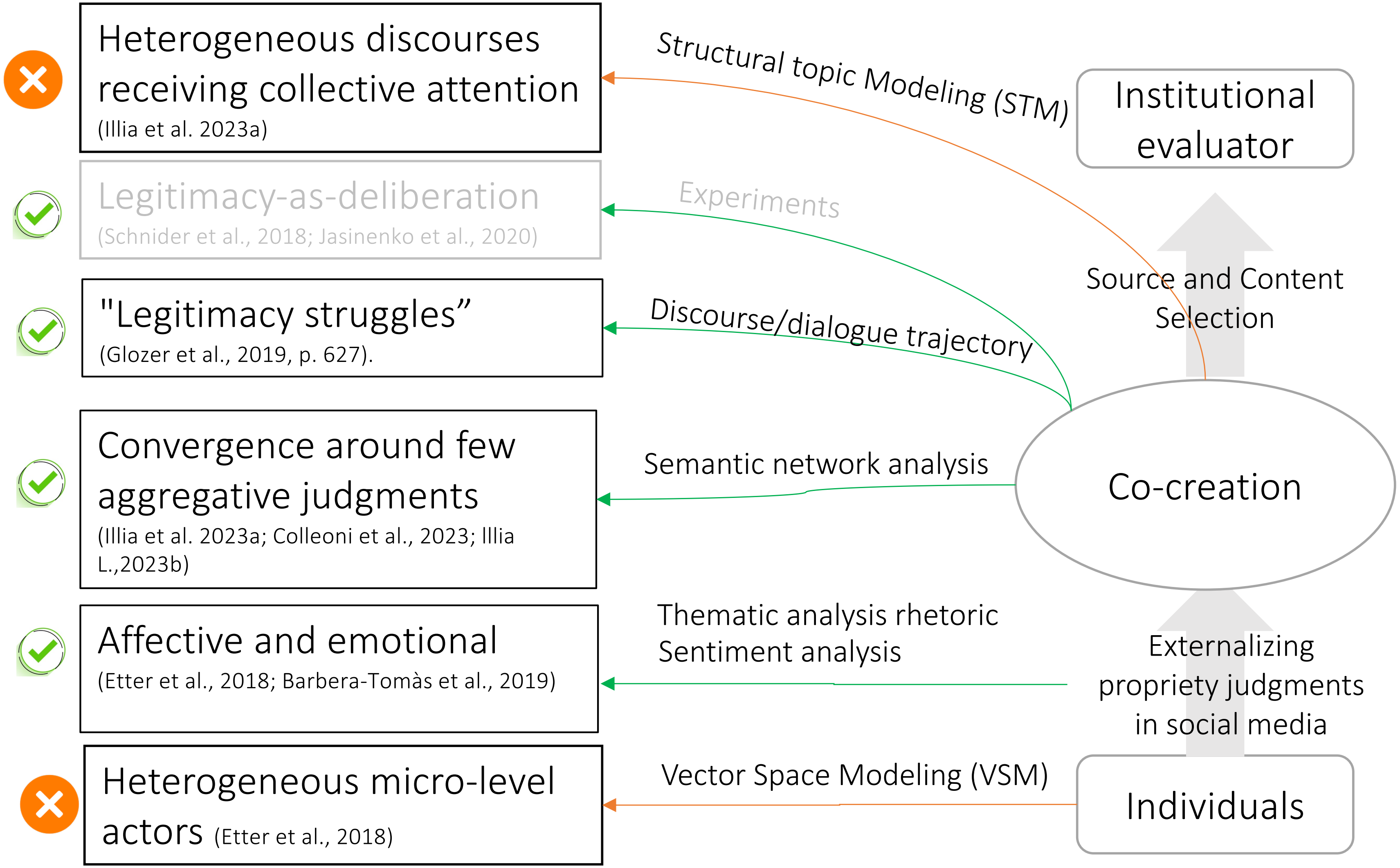
Long-term pick of attention



Temporary pick of attention

In definitive...

STM and VSM as promising but still to be validated as ideal for analyzing... this work is unpublished, yet (I put it in orange...)



This is us in this paper with STM and VSM...

References

- Alemany Oliver, M. & Vayre, J.-S. 2015. Big data and the future of knowledge production in marketing research: Ethics, digital traces, and abductive reasoning. *Journal of Marketing Analytics*, 3(1): 5-13.
- Aranda, A. M., Sele, K., Etchanchu, H., Guyt, J. Y., & Vaara, E. (2021). From Big Data to Rich Theory: Integrating Critical Discourse Analysis with Structural Topic Modeling. *European Management Review*, *18*(3), 197-214. https://doi.org/10.1111/emre.12452
- Baumer, E. P., Mimno, D., Guha, S., Quan, E., & Gay, G. K. 2017. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology*, 68(6): 1397-1410.
- Colleoni, E., Zyglidopolous, S., Illia , L. (2023) Stakeholders influence on firm behavior in the digital age. (2023). In *Academy of Management Perspectives*.
- Etter, M., Colleoni, E., Illia, L., Meggiorin, K., & D'Eugenio, A. (2018). Measuring Organizational Legitimacy in Social Media: Assessing Citizens' Judgments With Sentiment Analysis. (2018). In *Business & Society*https://doi.org/10.1177/0007650316683926
- Evans, J. A., & Aceves, P. (2016). Machine Translation: Mining Text for Social Theory. https://doi.org/10.1146/annurev-soc-081715-074206
- Glozer, S., Caruana, R., & Hibbert, S. A. (2018). The Never-Ending Story: Discursive Legitimation in Social Media Dialogue. *Organization Studies*. https://doi.org/10.1177/0170840617751006
- Jasinenko, A, Bàchtiger, A., Haack, P. (2022) The Role of Consensus in Legitimation through Deliberation: An Experimental Approach. Academy of Management Proceedings
- Haack, P., Schilke, O., & Zucker, L. (2021). Legitimacy Revisited: Disentangling Propriety, Validity, and Consensus. *Journal of Management Studies*, *58*(3), 749-781. https://doi.org/10.1111/joms.12615
- Illia, L., Etter, M., Meggiorin, K. and Colleoni, E. (2022), "From Micro-Level to Macro-Level Legitimacy: Exploring How Judgments in Social Media Create Thematic Broadness at Meso-Level", . (2022). In *Gegenhuber, T., Logue, D., Hinings, C.R.(B). and Barrett, M. (Ed.) Digital Transformation and Institutional Theory (Research in the Sociology of Organizations, Vol. 83), Emerald Publishing Limited, Bingley, pp. 111-131*.
- lllia. L., Colleoni, E., Etter, M., Meggiorin, K. (2023a) Finding the tipping point: when heterogeneous evaluations in social media converge and influence organizational legitimacy. Business and Society 62(1): 117-150. (2023).
- lllia L., Colleoni, E., (2023) Digital Corporate Communication & Issues Management: Monitoring digital collectives and tipping points of refraction around issues. (2023). In *Luhoma V., and Badham, M. (Eds) HANDBOOK OF DIGITAL CORPORATE COMMUNICATION Edward Elgar Publishing*.
- Martin W Bauer, Ahmet Süerdem, Aude Bicquelet. (2014) Text analysis: an introductory manifesto SAGE Benchmarks in Social Research
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
- Schnider R., Haack P., Scherer A. G. (2018). Legitimacy evaluation formation as deliberation: Evidence from experiments on corporate tax avoidance. *Academy of Management Proceedings*
- Roberts, M. E., Stewart, B. M., & Tingley, D. 2016. Navigating the local modes of big data. Computational social science, 51., 91: 1-40.
- Roberts, M. E., Stewart, B. M., & Tingley, D. 2019. Stm: An R package for structural topic models. Journal of Statistical Software, 91: 1-40.