A Computational Approach to Modeling Online Identity Discourses

Adam Papendieck, The University of Texas at Austin, apapendieck@utexas.edu

Abstract: The contemporary participatory media ecologies within which connected learners work and change themselves may be only loosely organized by disciplinary norms and categories. In such loosely disciplined contexts, learner identity, traditionally understood as an object of disciplinary power from a standard sociocultural (CoP) perspective, may be better understood as an ongoing dialogic project of discursive positioning. This study demonstrates a computational approach to modeling identity in an open, online network of ed-tech innovators. Latent Dirichlet Allocation (LDA) is combined with post hoc qualitative analysis to represent and make sense of online identity profiles as composites of seven latent network discourses. The study operationalizes useful theoretical notions of polyphony and positioning for the study of connected learning and identity in loosely disciplined participatory networks.

Introduction and conceptual framework

This study was motivated by the goal of developing new ways of making sense of and designing to support complex, evolving configurations of learners in open, networked environments. Models are one way to organize and represent complex phenomena for inquiry and action. Here I describe a computational approach to modeling identity discourses in an online network of ed-tech innovators, an approach that helps represent and treat learner identity as polyphonic, situated and dynamic.

From a sociological perspective, contemporary connected learners and networked knowledge workers might conveniently be theorized as *entrepreneurs of the self* (Foucault, 2008), crafting and re-crafting their subjective identities to differentiate themselves per the pervasive power of the neoliberal market rather than (or in addition to) the restrictive disciplinary power of schools or institutions (McNay, 2009). While standard ethnographic methods have been useful in elaborating the dynamics of identity construction in disciplinary contexts and communities of practice, the engagements of connected learners in online participatory netwroks tend to be highly mediated by technology, distributed, discontinuous and poorly-bounded by disciplines. As social learning environments, they may not generate or enforce disciplinary roles or recognizable "identity kits" (Gee, 1989, p. 7) in strong or persistent ways, and may not conform to traditional sociocultural understandings of community, cyclical practice or domain (Engeström, 2007). In such *undisciplined* contexts, where the disciplinary or community "center does not hold" (Engeström et al., 1999), we might understand identity less as a persistent psychological entity or an overarching project rooted in a disciplinary context, and more as an ongoing entrepreneurial and creative act of discursive positioning in a heterogeneous cultural, technological and practical landscape. From a Bakhtinian point of view, our subjective identity discourses—the stories we tell about ourselves to others—should be understood as situated and crafted *dialogically* in particular historical and social contexts, inevitably echoing and responding to other subjective discourses. Identity discourses are *polyphonic* constructions (Bakhtin, 1984). Identities are also positional acts: we may craft our identity discourses on the fly for specific reasons in different situations, allowing us to position ourselves in terms of status, power, knowledge, and sanctioned performances (Deppermann, 2015).

This study demonstrates how Latent Dirichlet Allocation (LDA) (Blei et al., 2003) can be used to infer in a systematic, probabilistic manner a set of identity discourses from a relatively large collection of online profiles, in this case, those created by members of an open, urban ed-tech innovation network. As a type of computer-assisted narrative analysis, this approach presents itself as a way to infer different discourses that play into individual stories of identity on the network, and make them available for deeper analysis of how they form around and pattern network positioning and learning interactions.

Method

The work carried out in this study is part of a broader mixed methods case study of the composition and enactment of the Ed-tech Network (pseudonym), a case of an open, urban, digitally-mediated innovation network. The Ed-tech Network assembles online via the meetup.com social networking platform as well as at TechAssembly (pseudonym), a co-working space and startup incubator in downtown Neotown (pseudonym). All 2,241 public "ed-techer" profiles posted to the network's meetup.com group by the start of this study were accessed via public API and transformed into a corpus for topic modeling using the 'tm' package in R (Feinerer, 2018). This transformation involved routine natural language processing steps, including stopword removal (per the default

English "en" list in the 'tm' package) and word stemming (e.g. truncating words like "science" and "sciences" to a common stem: "scienc").

LDA is a common computational method for identifying latent structure in text corpora, having been employed to model topics within and among large documents like articles and online reviews (e.g. Evans, 2014; Nichols, 2014), as well as corpora of documents smaller than the typical ed-techer profile, like sub-120 character tweets (Jónsson & Stolee, 2015). LDA is a probabilistic and generative model; it uses iterative, Bayesian word sampling from documents in a corpus to automatically detect topics based on the likelihood of term co-occurrence (Blei et al., 2003). The approach infers through a series of model runs a probable set of topics (collections of terms) that generate the documents. Topic probabilities can also be assigned to each document in an a posterior manner, representing individual documents based on term probabilities. In the way that LDA can be used to simultaneously *surface topics in a corpus* and also *represent individual documents in the corpus as composites of those topics*, and the modeling approach can be thought of as a computational way of operationalizing the Bakhtinian notion of a polyphony (Bakhtin, 1984) for the analysis of emergent network discourse. Individual stories on the network can be modeled (understood, represented and organized) as emergent, polyphonic mixtures of topics, topics that are themselves inferred through the iterative, probabilistic allocation of words from stories told by individual voices in the network.

In order to fit an LDA model to a corpus, the number of topics (k) must be fixed a priori. While there are methods of "cross-validating" the number of topics based on harmonic means and perplexity minimization (Blei et al., 2003; Grün & Hornik, 2011; Ramage & Rosen, 2009), a primary research consideration is the hermeneutic utility of the number and composition of the topics themselves. Models specifying very large numbers of topics (e.g. $k > 10$) may become nearly as inscrutable as the aggregate of individual texts from which the topics are inferred. For the purposes of this study, I was interested in producing a small, intuitive set of topics for interpreting individual stories. To develop this set, I took a systematic approach to the qualitative evaluation of the hermeneutic utility of topics modeled per different a priori numbers of topics (k). This approach, which I call *term tracing*, involved conducting iterative model runs to predict from 2 to 10 topics using the 'topicmodels' package in R (Grün & Hornik, 2011), and comparing topics generated at each iteration in a stepwise progression, proceeding from the lowest to the highest k model run. The last topic included in the hermeneutically useful set presented here was the last to persist recognizably across two sequential model runs (k and $k + 1$) based on its term composition. These *recognizably persistent topics* across model runs are adopted for the purposes of this study as *fundamental discursive topics,* or what I call *latent discourses.*

For each latent discourse, I then used the 'wordcloud' package (Fellows, 2018) in R to create a frequency word cloud of the top 20 terms generated across all model runs for which the topic was recognizable. This provides a summary at a glance of the high probability terms that tend to persist across model runs, that is, to be regularly assigned to each recognizable topic across model runs. I then examined the set of 20 ed-techer profile documents assigned to each topic with the highest degree of probability at the highest k run for which all topics are present, and synthesized a simplified, abstracted version of each profile to show how latent discourses play out in their network stories. The overall analysis amounts to a systematic, computational, probabilistic way of identifying and qualitatively interpreting different discourses at play in a relatively large dialogic storytelling network.

Findings

Tracing terms across LDA model runs for 2 to 10 topics (k), I identified seven persistently recognizable identity discourses based on meetup profile responses to the prompt "tell us a bit about yourself." I will examine these identity discourses in the order in which they were inferred in the stepwise progression of LDA model runs.

The first latent discourse to emerge is what I call a *teacher* identity discourse (Figure 1A). Based on post hoc analysis of the 20 profiles assigned to this topic with the highest degree of probability, I found that almost all ed-techers using this discourse reported being either a former teacher (13/20 sampled), current teacher (3/20) or preservice teacher (1/20). Of the 13 identifying as former teachers, nearly a third (4) reported being former Teach for America (TFA) teachers. Overall about half (9/20) also reported prospective, current or recent university student status, suggesting that ed-techer teacher identities may be newly formed or forming. The remaining ed-techers using the teacher identity discourse included a school robotics coach (possibly a volunteer), an instructor in a post-secondary for-profit academy and a school-community liaison.

The next discourse inferred through LDA is represented by terms like "business," "software," "develop," "manage" and "profession" (Figure 1B). Upon inspection, these profile stories position ed-techers in business and tech industry roles, like UX developer, project manager, ed-tech business development specialist, marketer and software and web developer. The discourse is used by ed-techers to identify themselves in terms of roles and divisions of labor in the context of commercial firms. I call this a *commercial professional* identity discourse.

The next discourse to surface (Figure 1C) is distinctive in the way it is employed chronotopically: ed-techers present themselves in states of motion or stasis through space and time, that is, in terms of where they have been, where they are, and where they are going in terms of their work context. Many of these people story themselves as recent transplants to the city. Some are new to town and looking for a job, while others want to get involved with and start new things, like hackathons. Many also express a passion not just for ed-tech but for the broader ecosystem and cultural "vibe" of Neotown, including its "music," food and "sunshine" (meetup profiles). Ed-techers identifying as longtime Neotown residents employ this discourse to cast themselves as having grown up in, become fans of, or fallen in love with the city. This is a discourse that patterns identity stories in terms of cultural, professional, temporal and geographic transition. It is a chronotopic mobility discourse, and it is used by ed-techers to identify as new, old, local or from abroad. I call this the *mobile* identity discourse.

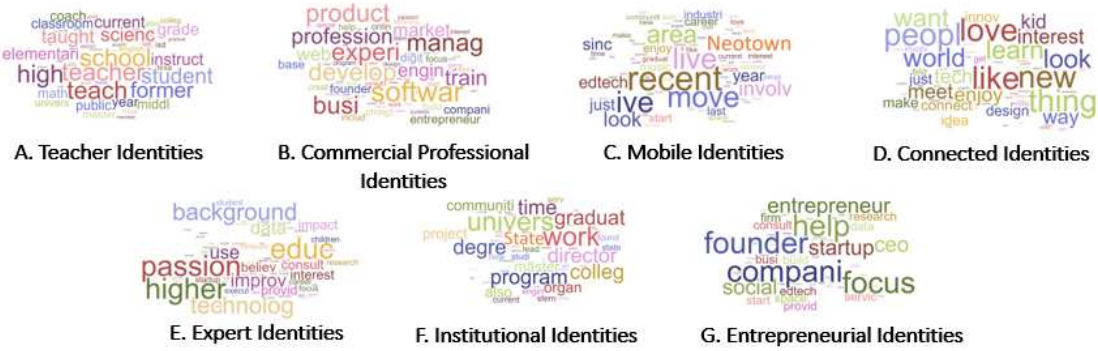

Figure 1. Latent identity discourses inferred via LDA that persisted recognizably across all model runs.

Ed-techers use the next discourse (Figure 1D) to present themselves as people who "love" to "meet" people, "connect" and "share" "new" and "innovative" things (meetup profiles). This *connected* identity discourse is used to emphasize ed-techer desires to be enmeshed as creators, sharers and learners in an innovative community. A longtime product manager, for instance, says they want to meet people to build stuff, change education, and host networking events. These ed-techers emphasize openness and sharing for learning and collaborative creation. Ed-techers from a variety of work contexts engage this connected identity discourse.

The next identity discourse (Figure 1E) helps ed-techers position themselves as experienced technology experts with a "passion" for education (meetup profile). The stories representing this discourse are diverse, but upon inspection, ed-techers engage what I call an *expert* discourse to emphasize historical dedication to and experience in the educational technology domain. Some emphasize that they are, for example, experienced in technology and startups and passionate about education, while others emphasize background in higher ed and k-12 technology. From private sector and university work contexts, these ed-techers write that they engage technology for educational change at different levels, from primary to "higher" education (meetup profiles).

The stories generated via the sixth identity discourse, again, reflect a diversity of professional backgrounds and positions (Figure 1F). Looking at the high probability terms, we can see it has something to do with word stems like "univers," "graduat," "colleg" and "work." Profile inspection reveals ed-techers using this discourse to elaborate their university affiliations, degrees, and work experience with programs and institutions. Ed-techers often use this discourse to describe themselves as qualified and experienced leaders, specifying recognizable institutional affiliations, typically universities. I call this a discourse of *institutional* identity.

Finally, we have what I will call an *entrepreneurial* identity discourse (Figure 2G). These ed-techers use the discourse to story themselves as designers, startup CEOs and founders of entrepreneurial for- and non-profits that "research" and "focus" on issues and markets, and "build" things that "help" and "empower" (meetup profiles). The "social" outcomes of entrepreneurial activity are often featured in this discourse. For instance, one ed-techer who describes himself as an alumnus of incubator and accelerator programs and a social entrepreneur is interested in making an impact on the education market and meeting demand among students for better learning and growth through education. As opposed to the managerial and technical roles described using the comercial professional identity discourse, entrepreneurial stories emphasize high-agency, independent, creative identities, like startup founders, designers, "freelancers," "free-agents" and "consultants" (meetup profiles).

Discussion and implications

Overall, I am able to recognize and make sense of seven modeled discourses about identity that ed-techers leverage to describe themselves upon joining the meetup: *teacher, commercial professional, mobile, connected, expert, institutional* and *entrepreneurial*. It is critical to note that LDA represents each individual meetup profile story as

a composite of multiple latent identity discourses available to ed-techers and operating in the network. Individuals can be distinguished and characterized quantitatively based on the relative probability that their identity story is patterned by some discourses more than others. LDA, therefore, can be used as a technical, computational way of modeling ed-techer identity stories as polyphonic (Bakhtin, 1984): they are comprised of and speak back to a variety of latent discourses on the network. One ed-techer, for instance, invokes both a teacher identity and institutional identity as *a former public school teacher, a current recruiter for a charter school and a current MBA student at State U. Neotown.* (abstracted text of meetup profile). Another combines mobile, teacher and institutional identities, positioning herself as *new to town, a former TFA computer teacher, recent MEd* (abstracted text of meetup profile). Yet another combines commercial professional and connected identities as *a long-time product manager interested in meeting people to build stuff, change education, and host networking events* (abstracted text of meetup profile). Conceptualizing ed-techers as entrepreneurs of the self, these profile stories might be read as *entrepreneurial narratives of the self* that help ed-techers position and market themselves on a dynamic and *undisciplined* participatory media landscape. Going forward, the analysis of learner *identity* presented here will be integrated with a similar examination of latent discourses about learner *interests* inferred from another section of the ed-techer profile. This will enable a systematic characterization of how ed-techer identities relate to the multitude of (potentially very different) interests, goals and objectives around which the Ed-tech Network assembles.

While standard sociocultural models have proven useful for understanding and designing for learning and identity development in communities of practice and disciplines, the analysis presented here advances the field of connected learning by demonstrating how notions of *polyphony* (Bakhtin, 1984) and *positioning* (Deppermann, 2015) can be operationalized to model learner identities as emergent composites that speak within, across and through a variety connected discourses and communities. Empirically-informed sets of identity and interest parameters like those discussed here could be used to structure deeper mixed methods, quantitative or qualitative analyses of connected learning dynamics in loosely disciplined networks. Such parameters could also be used to develop composite "user profiles" for the intentional design of more equitable, just and effective open learning and innovation networks.

References

Bakhtin, M. M. (1984). *Problems of Dostoevsky's Poetics* (C. Emerson, Trans.; Vol. 8). University of Minnesota Press.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Deppermann, A. (2015). Positioning. In A. De Fina & A. Georgakopoulou (Eds.), *The Handbook of Narrative Analysis* (pp. 369–387). John Wiley & Sons, Inc.

Engeström, Y. (2007). From communities of practice to mycorrhizae. In J. Hughes, N. Jewson, & L. Unwin (Eds.), *Communities of practice: Critical perspectives* (pp. 41–54). Routledge.

Engeström, Y., Engeström, R., & Vähäaho, T. (1999). When the center does not hold: The importance of knotworking. In S. Chaiklin, M. Hedegaard, & U. J. Jensen (Eds.), *Activity theory and social practice: Cultural-historical approaches* (pp. 345–374). Aarhus University Press.

Evans, M. S. (2014). A Computational Approach to Qualitative Analysis in Large Textual Datasets. *PLOS ONE*, *9*(2), e87908.

Feinerer, I. (2018). *Introduction to the tm Package Text Mining in R*. http://cran.uib.no/web/packages/tm/vignettes/tm.pdf

Fellows, I. (2018). *Package 'wordcloud.'* https://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf

Foucault, M. (2008). *The birth of biopolitics: Lectures at the Collège de France, 1978-1979* (G. Burchell, Trans.). Palgrave Macmillan.

Gee, J. P. (1989). Literacy, Discourse, and Linguistics: Introduction. *Journal of Education*, *171*(1), 5–17.

Grün, B., & Hornik, K. (2011). topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software*, *40*(13).

Jónsson, E., & Stolee, J. (2015). *An Evaluation of Topic Modelling Techniques for Twitter*. http://www.cs.toronto.edu/~jstolee/projects/topic.pdf

McNay, L. (2009). Self as Enterprise: Dilemmas of Control and Resistance in Foucault's The Birth of Biopolitics. *Theory, Culture & Society*, *26*(6), 55–77.

Nichols, L. G. (2014). A topic model approach to measuring interdisciplinarity at the National Science Foundation. *Scientometrics*, *100*(3), 741–754.

Ramage, D., & Rosen, E. (2009). *Stanford Topic Modeling Toolbox*. Stanford Topic Modeling Toolbox. https://nlp.stanford.edu/software/tmt/tmt-0.4/