

Context collapse and identity complexity in online self-presentation: a cross-platform comparison [★]

Samantha C. Phillips¹ and Kathleen M. Carley¹

CASOS Center, Software and Societal Systems Department
Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213
{samanthp, carley}@andrew.cmu.edu

Abstract. Social media users curate their self-presentation for audiences that vary in breadth and predictability across platforms. When users are uncertain about who will encounter their profile and what norms those viewers will bring, they may limit identity complexity in their self-presentation, defined as the number of identities expressed and the conceptual and cultural diversity among them. To assess this possibility, we compare profiles from X and Bluesky users who posted about Eurovision during the same period in May 2026. X predates Bluesky by about 16 years, has a user base several times larger, and defaults to an algorithmically curated feed that exposes users to content from beyond those they follow, contributing to greater audience uncertainty on X. We find that across all three dimensions, Bluesky users expressed greater identity complexity than X users. Specifically, they presented more identities, more conceptually diverse identities, and more culturally atypical combinations of identities, although differences were small overall. These findings are consistent with the possibility that broader and less predictable audiences constrain complex self-presentation and suggest that platform context can shape the forms of identity expression that users make visible online.

Keywords: Self-presentation · Identity complexity · Context collapse · Invisible audience · Social media.

1 Introduction

Social media connects people across geographic and cultural boundaries, exposing users to audiences that are often broader and less visible than those

[★] This work was supported in part by the Knight Foundation and the Office of Naval Research grant MURI: Persuasion, Identity, & Morality in Social-Cyber Environments, N00014-21-12749. Additional support was provided by the Center for Computational Analysis of Social and Organizational Systems (CASOS) and Center for Informed Democracy and Social-cybersecurity (IDeaS) at Carnegie Mellon University. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Knight Foundation, Office of Naval Research, or the U.S. Government.

encountered in offline settings. People and expectations that would ordinarily remain separated across social contexts can converge, a phenomenon commonly referred to as context collapse. Because users rarely know exactly who will encounter their profiles, they must construct an imagined audience in the face of substantial audience uncertainty [8, 6]. This uncertainty may therefore shape not only what people choose to disclose, but how they organize and present different aspects of who they are. This matters because the identities people make visible influence how others perceive them and can also feed back into how they understand themselves.

Broader, more diverse and potentially unknown audiences create competing pressures for self-presentation. On the one hand, there are more opportunities for social connection. Users may therefore benefit from displaying multiple facets of themselves because each provides another possible point of recognition. On the other hand, broader and less visible audiences are more difficult to anticipate. Greater audience uncertainty increases the risk that signals appropriate for one audience will be misunderstood or negatively evaluated by another. Work on context collapse therefore suggests an opposing incentive to present oneself in ways that remain intelligible across audience segments [8]. For example, when multiple audiences are collapsed into a single context, users may constrain their self-presentation toward a “lowest common denominator” – signals likely to be acceptable across the audiences they anticipate encountering [5] – or place disproportionate weight on the expectations of a particularly consequential “strongest audience” [7].

We examine these possibilities through expressed identity complexity, or the structure of the identities users present in their profiles. We use identities to refer to socially meaningful, self-relevant concepts through which people understand themselves and are recognized by others, such as occupations, family roles, political affiliations, nationalities, and hobbies [14]. We distinguish three dimensions of expressed identity complexity: the number of identities presented, their conceptual diversity (i.e., semantic dissimilarity), and their cultural atypicality (i.e., how unusual it is for those identities to be expressed together by the same person).

Previous research on context collapse and self-presentation focuses on what users select to disclose or censor, and how they manage privacy settings [4, 13]. Furthermore, these effects are often examined on one platform at a time. Yet context collapse is not a uniform property of social media [1]. The breadth, visibility, and predictability of one’s potential audience depend partly on the architecture and population of the platform on which self-presentation occurs. We extend these accounts by examining how these pressures alter the structure of identities people publicly claim in their profile on two social media platforms: X (formerly Twitter) and Bluesky.

X and Bluesky are a particularly useful comparison because Bluesky emerged from a project initially developed within Twitter and retains many of the same basic conventions of public, real-time discourse, while developing a different architecture, history, and user population [11]. They have several key distinctions.

First, they differ in the amount and structure of information users can present in their profile. Bluesky permits a longer profile description than X and the platforms provide somewhat different auxiliary profile fields, such as a “pronouns” field on Bluesky and “location” field on X. Greater expressive capacity provides a straightforward reason to expect more identities to be made visible on Bluesky. Second, there are reasons to believe that these platforms differ in audiences and norms. X is substantially older and larger, and its default algorithmic feed routinely exposes users to content and users beyond those they explicitly follow [3]. Bluesky places greater emphasis on following-based and user-curated feeds [11, 2]. There is also strong evidence that Bluesky’s audience is particularly politically homogeneous: recent work finds that its user base is predominantly left-leaning and that opposing political positions are often represented by only a small minority of users [11, 12].

These platform differences do not provide a causal test of context collapse, but they make it reasonable to expect that profiles on X circulate before a broader and less predictable potential audience. If broader audiences primarily increase opportunities for social connection, X users might display identities spanning a wider range of conceptual and cultural domains. If audience uncertainty instead favors broadly legible self-presentation, as the context collapse account suggests, identities displayed on X should be less conceptually diverse and less culturally atypical than those displayed on Bluesky.

We evaluate these expectations by comparing identity complexity in profiles from X and Bluesky users collected over the same 32-day period in May 2026 using the same Eurovision-related keywords. Across both profile bios and the full set of available profile fields, Bluesky users express more identities, and those identities are more conceptually diverse and culturally atypical than those expressed on X. This pattern is consistent with the expectation that broader and less predictable audiences are associated with more constrained forms of identity presentation, reinforcing previous findings that context collapse can reduce variability in self-presentation and encourage more broadly acceptable forms of expression.

2 Methods

2.1 Data

This work uses X (formerly Twitter) and Bluesky data collected using a keyword search between April 30, 2026 through May 31, 2026. First, we obtained Bluesky data from the Indiana University Observatory on Social Media (OSoMe) AT Protocol archive¹. We identified posts containing Eurovision-related keywords and extracted the users associated with those posts, including users connected through replies, quotes, and reposts. We then collected profile information for the resulting set of unique users using the public Bluesky AppView API, so we

¹ <https://osome.iu.edu/resources/data-resources>

have profile information for each user at one point in time. This process yielded profile information for 74,215 unique Bluesky users.

Second, we collected X data using the official X API and the same Eurovision-related keywords used for Bluesky. Because API access was usage-priced, collection proceeded sequentially by day under a fixed overall budget, with larger daily allocations for the Eurovision semi-finals and final. Posts were retrieved in pages of up to 100. After each page, we identified unique authors, removed users whose profiles had already been collected on an earlier day, and recalculated the remaining budget while reserving sufficient funds to retrieve profile metadata for newly observed users. We continued retrieving posts until another page could not be safely accommodated within the day’s allocation. Profile information was then collected once for each newly observed user, yielding 48,790 unique X users.

2.2 Measuring identity complexity

We first segmented bios into candidate identity phrases using a rule-based extraction pipeline designed for bios [9]. Specifically, we normalized Unicode, punctuation, whitespace, hashtags, emojis, pronouns, ages, locations, and list-like constructions, while removing URLs, emails, and phone numbers. We reconciled the phrases from this process with an already cleaned set from a dataset of 31.6M X bios [10]. To improve coverage, we then re-scanned the original bio text for all phrases in this dictionary, matching longer phrases first. Our retained identities are primarily English-language phrases or emojis, meaning that the procedure does not capture all forms of identity expression. In addition, some users express highly idiosyncratic identities for which co-occurrence patterns cannot be estimated reliably. Overall, we retained at least one identity for 43,074 Bluesky bios and 19,692 X bios.

We operationalized three dimensions of identity complexity. The first dimension is the number of cleaned identities retained from a bio. To measure conceptual diversity and cultural atypicality among identities, we first identified all unique pairs of identities within each bio. For conceptual diversity, we embedded each cleaned identity using the all-mpnet-base-v2 sentence-transformers model² and calculated cosine similarity for every pair of concepts. For cultural atypicality, we used normalized pointwise mutual information (NPMI) values estimated from identity co-occurrence patterns using the larger corpus of X bios [10]. As a validation of this reference corpus, pairwise identity co-occurrence frequencies were moderately correlated with those observed in the present X/Bluesky sample among pairs observed in both datasets (Spearman’s rank correlation coefficient=0.55; N=626,189 pairs).

At the bio level, conceptual diversity was calculated as one minus the average normalized semantic similarity across all identity pairs, and cultural atypicality was calculated as one minus the average normalized co-occurrence association. Thus, higher values indicate greater complexity across all three dimensions: more identities, identities that are more semantically diverse, and identities that are

² <https://huggingface.co/sentence-transformers/all-mpnet-base-v2>

more rarely co-expressed. Spearman’s correlation between the number of identities and conceptual diversity is 0.006; between number of identities and cultural atypicality is 0.033; and between conceptual diversity and cultural atypicality is 0.33.

2.3 Analysis

We compare identity complexity across platforms using Welch’s two-sample t-tests, which do not assume equal variances between groups. We estimate separate tests for the number of identities, conceptual diversity, and cultural atypicality, and report Cohen’s d to characterize the magnitude of each platform difference. Our primary analyses are restricted to non-empty bios and measure identities using bio text only. As robustness checks, we repeat these analyses but include the empty bios and identities extracted from all available profile fields, rather than just the bio. We also estimate OLS models predicting conceptual diversity and cultural atypicality as a function of platform and the number of identities expressed.

3 Results

First, we summarize identities presented in bios in each sample. A higher proportion of bios in our Bluesky sample were non-empty (94.0%) than in our X sample (77.7%). Among non-empty bios, Bluesky users presented an average of 2.15 (std. dev. = 2.98) identities, compared with 1.53 (std. dev. = 2.37) identities among X users. The median number of identities was 1 on both platforms. Notably, when all available profile fields are included, the average number of identities increases to 2.41 (std. dev. = 3.12) on Bluesky and 2.26 (std. dev. = 2.77) on X.

Second, we assess differences in each dimension across platforms. Across all three dimensions, identity complexity was higher on Bluesky than on X. Conceptual diversity was slightly higher on Bluesky (mean = 0.762, std. dev. = 0.060) than on X (mean = 0.756, std. dev. = 0.075), while cultural atypicality was 0.918 (std. dev. = 0.061) on Bluesky and 0.898 (std. dev. = 0.093) on X. The largest differences were observed for cultural atypicality ($d = 0.28$) and the number of identities expressed ($d = 0.23$), whereas the difference in conceptual diversity was substantially smaller ($d = 0.10$). Thus, although the differences were consistent in direction, their magnitudes were small overall, particularly for conceptual diversity.

These differences were robust to including empty profiles, measuring identities across all available profile fields, and, for the two relational dimensions, accounting for the number of identities expressed. In fact, controlling for the number of identities explained almost none of the platform differences in these dimensions: the Bluesky–X difference in conceptual diversity changed only from 0.0064 to 0.0062, and the difference in cultural atypicality changed from 0.0203 to 0.0193.

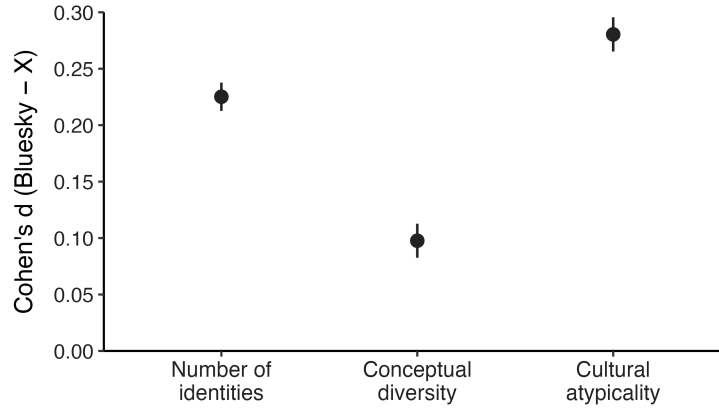


Fig. 1. Cohen's d effect sizes comparing Bluesky with X/Twitter across the three complexity dimensions. Points represent standardized mean differences (Bluesky - X); vertical lines show 95% confidence intervals. Positive values indicate higher average scores on Bluesky.

4 Discussion

Our cross-platform comparison revealed a consistent, though modest, pattern: identity presentation was more complex on Bluesky than on X. Specifically, Bluesky users expressed more identities, and those identities were more conceptually diverse and more culturally atypical than those expressed on X. This pattern is compatible with the possibility that broader and less predictable audiences constrain complex forms of self-presentation. X exposes users to a larger, more heterogeneous, and more algorithmically mediated audience, creating greater uncertainty about who may encounter a profile. Such context collapse may encourage users to limit the identities they present and to combine them in more legible and culturally conventional ways. By contrast, the comparatively narrower and more like-minded audience on Bluesky may reduce these pressures, allowing users to express a greater number and range of identities.

Differences in self-presentation affordances may also help explain part of this pattern. In particular, the greater space and profile flexibility available on Bluesky may help explain why users expressed more identities. The differences in conceptual diversity and cultural atypicality, however, extend beyond the amount of identity information users provide, suggesting that the platform differences are not limited to expressive capacity alone.

This work is not without limitations. First, we did not attempt to connect users across platforms. To mitigate temporal and topical differences, we collected users on X and Bluesky using the same keywords, during the same period. Future work should extend these analyses to examine the same users on multiple platforms. Second, we only discussed a small set of theoretically motivated differences between platforms. X and Bluesky differ on many dimensions beyond those

considered here, including institutional anchoring and other affordances; future work should isolate these dimensions experimentally. On Bluesky, handles do not merely reflect the user-specific username. Rather, it also includes a domain component which allows users to refer to an organization or affiliation directly in their username. In this way, institutional affiliation may be more salient on Bluesky than X, where users do not attach these domains to their username. Other possible differences may be more relevant to within-platform contexts, such as specificity and ephemerality. Both platforms are used for real-time discussion, although X has a much longer-established role in breaking news and live events. Different discussions on these platforms may shift more quickly than others. Future work should expand on these dimensions of context considered and possibly adjudicate between their effects experimentally.

Finally, and most crucially, this work is entirely observational and between-person. Therefore, while we can claim the results are *consistent* with our theory, this does not establish causal effects. Future work should therefore move beyond cross-sectional platform comparisons to directly test the mechanisms proposed here. One approach would be to experimentally manipulate features of the imagined audience (e.g., its size, ideological or social heterogeneity, or the predictability of who will view a profile) and examine how participants alter their self-presentations in response. Researchers could similarly manipulate profile affordances, such as the amount of space available or the availability of dedicated fields for affiliations and other identity information, to isolate their effects on the number and types of identities expressed. Within-person studies could complement these experiments by comparing how the same users present themselves across platforms or across contexts with different audiences, reducing concerns that the observed differences primarily reflect platform selection. Qualitative interviews could also inform how users understand these contextual features and which audiences and risks they consider when deciding what identity information to disclose. Together, these approaches could help distinguish the effects of audience structure and platform affordances from compositional differences between platforms and provide stronger evidence for how social context shapes the structure of self-presentation.

References

1. Davis, J.L., Jurgenson, N.: Context collapse: Theorizing context collusions and collisions. *Information, communication & society* **17**(4), 476–485 (2014)
2. Failla, A., Rossetti, G.: “i’m in the bluesky tonight”: insights from a year worth of social data. *PloS one* **19**(11), e0310330 (2024)
3. Gauthier, G., Hodler, R., Widmer, P., Zhuravskaya, E.: The political effects of x’s feed algorithm. *Nature* **652**(8109), 416–423 (2026)
4. Gil-Lopez, T., Shen, C., Benefield, G.A., Palomares, N.A., Kosinski, M., Stillwell, D.: One size fits all: Context collapse, self-presentation strategies and language styles on facebook. *Journal of Computer-Mediated Communication* **23**(3), 127–145 (2018)

5. Hogan, B.: The presentation of self in the age of social media: Distinguishing performances and exhibitions online. *Bulletin of science, technology & society* **30**(6), 377–386 (2010)
6. Litt, E.: Knock, knock. who’s there? the imagined audience. *Journal of broadcasting & electronic media* **56**(3), 330–345 (2012)
7. Marder, B., Joinson, A., Shankar, A., Thirlaway, K.: Strength matters: Self-presentation to the strongest audience rather than lowest common denominator when faced with multiple audiences in social network sites. *Computers in Human Behavior* **61**, 56–62 (2016)
8. Marwick, A.E., Boyd, D.: I tweet honestly, i tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society* **13**(1), 114–133 (2011)
9. Pathak, A., Madani, N., Joseph, K.: A method to analyze multiple social identities in twitter bios. *Proceedings of the ACM on human-computer interaction* **5**(CSCW2), 1–35 (2021)
10. Phillips, S.C., Carley, K.M.: An assessment of identity complexity in social media bios. In: *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS)* (2026), accepted
11. Quelle, D., Bovet, A.: Bluesky: Network topology, polarization, and algorithmic curation. *PloS one* **20**(2), e0318034 (2025)
12. Salloum, A., Quelle, D., Iannucci, L., Bovet, A., Kivelä, M.: Politics and polarization on bluesky. *arXiv preprint arXiv:2506.03443* (2025)
13. Vitak, J.: The impact of context collapse and privacy on social network site disclosures. *Journal of broadcasting & electronic media* **56**(4), 451–470 (2012)
14. White, H.C.: *Identity and control: How social formations emerge*. Princeton university press (2008)