Can simple social copying heuristics explain tag popularity in a collaborative tagging system?

Jared Lorince & Peter M. Todd

Adaptive Behavior and Cognition Lab

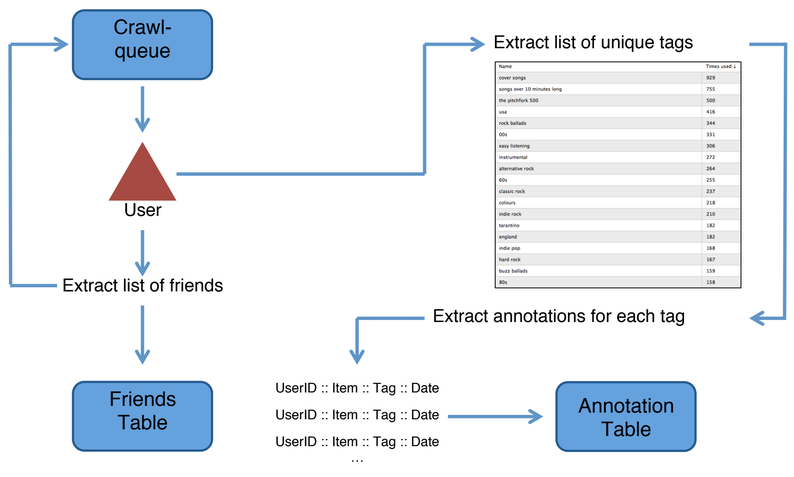
Abstract & contributions

We explore how simple models of boundedly rational human decision making may partly account for the emergent high-level properties of a complex social information environment, in particular the aggregate distribution of tag use (i.e. overall tag popularity) in a **collaborative tagging system**. Building upon previous work on tagging behavior [e.g. 1,2,3], our results indicate that **simple social copying mechanisms** can generate surprisingly good fits to the empirical data, with implications for the design and study of tagging systems. Our major contributions are:

- A novel methodology for crawling the social music website Last.fm;
- A set of possible social copying heuristic models in a tagging environment;
- Methods for estimating plausible parameter values for our heuristic models from empirical data; and,
- A set of multi-agent models employing social copying heuristics that demonstrate the extent to which the patterns of tag popularity we see in our crawled data can be explained through copying behavior.

Dataset

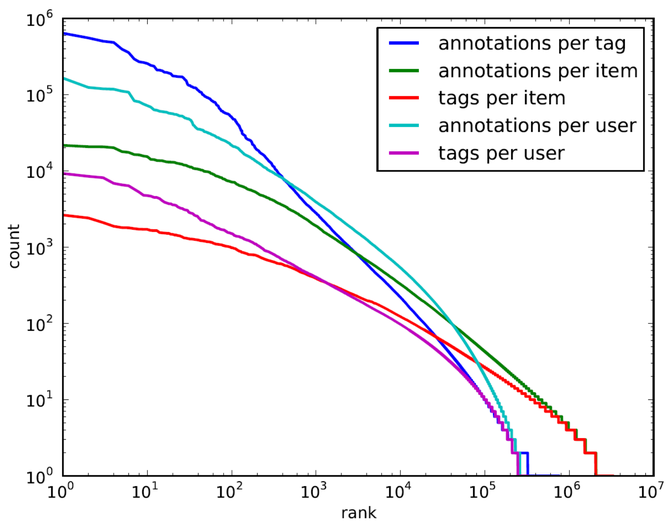
- Tagging data from music website Last.fm, a broad folksonomy [4] with freeform tagging
- Time-stamped annotation data from ~1 million users
- Crawled using a combination of API methods and HTML scraping of user profile pages,
- Total of ~33 million annotations across ~3 million artists, albums, and songs.



Schematic of crawling process.

Users	1,053,163
Active Taggers	318,415
Total Annotations	33,140,605
Total Unique Items	3,262,724
Total Unique Tags	747,275

Summary of data collected. “Active taggers” are those users who have tagged content at least once.



Rank-frequency plots of key metrics from the dataset, on a log-log scale.

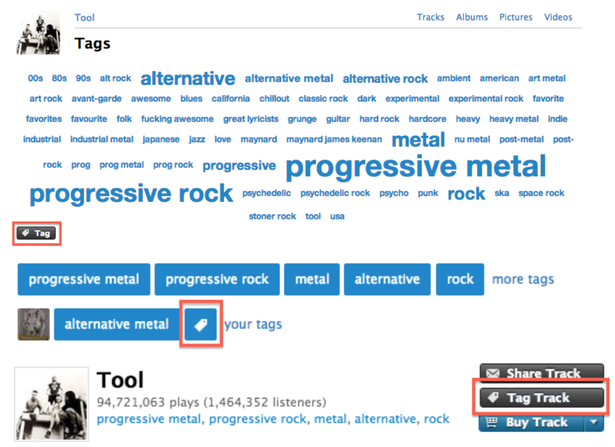
Social copying mechanisms

Complex choice environments with readily available social information are conducive to **social copying strategies** [5]. To what extent can simple, psychologically plausible heuristic copying strategies account for observed patterns of tag popularity? We proposed three simple strategies:

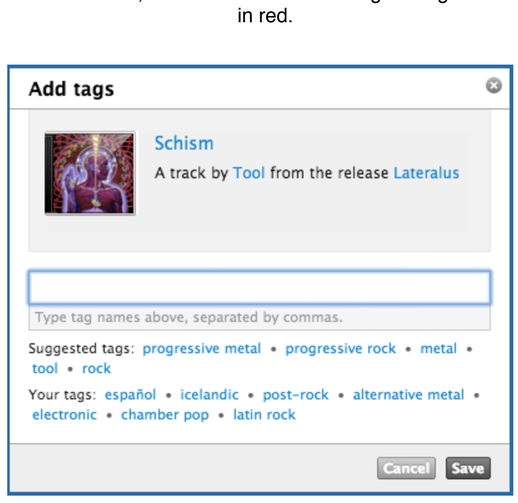
- **Uniform heuristic**: Select a tag randomly from the existing distribution for a given item
- **Normalized heuristic**: Copy an existing tag with probability proportional to its frequency in the tag distribution for that item
- **Top-five heuristic**: Randomly select one of the top five existing tags for an item.

Interface effects

The Last.fm interface structure facilitates use of the **top-5 heuristic**. Thus we hypothesized that it would perform best.



Three tag information views, with the icon for launching the tag interface highlighted in red.

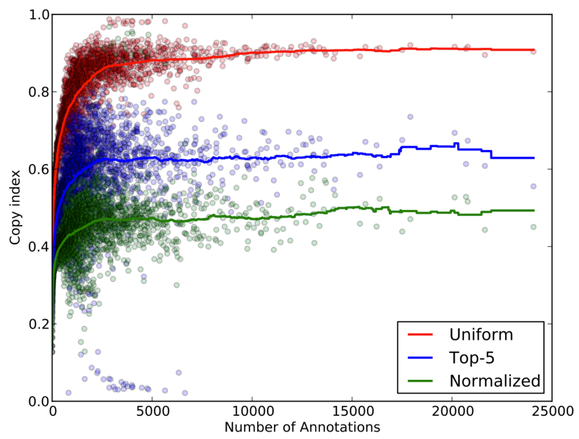


The tagging interface. “Suggested tags” are the top five most used tags for a given item.

Simulation framework

We developed **multi-agent simulations** with simple behavioral rules, and compared the emergent distributions of tag popularity to that observed in the empirical data.

- Mirrored attributes of empirical data: ~300,000 simulated agents generated ~33 million total annotations.
- Distributions of item popularity and user tagging levels matched empirical data
- No semantics (“tags” were arbitrary integers).
- Constrained the space of plausible parameter values by estimating copy indexes under each of the three heuristics (i.e. the mean proportion of annotations that were “copies” under our heuristic definitions).



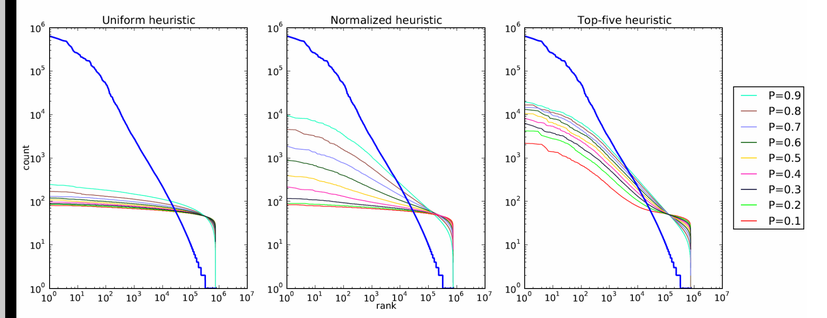
Copy index as a function of total annotations. Each point indicates the mean copy index across all months for items with the corresponding number of annotations. Solid lines are moving averages of increasing window size equal to $a^{0.9}$, where a is the total number of annotations shown on the x-axis.

Basic models

The three basic models had a single parameter, *P*, the probability of an agent engaging in copying in a particular instance of tagging an item. As expected, the top-5 model generated the best fits to the data.

Model Summary:

- With probability *P*, agents copy an existing tag for a particular item
- With probability *1-P* agents assign a random tag to the item.



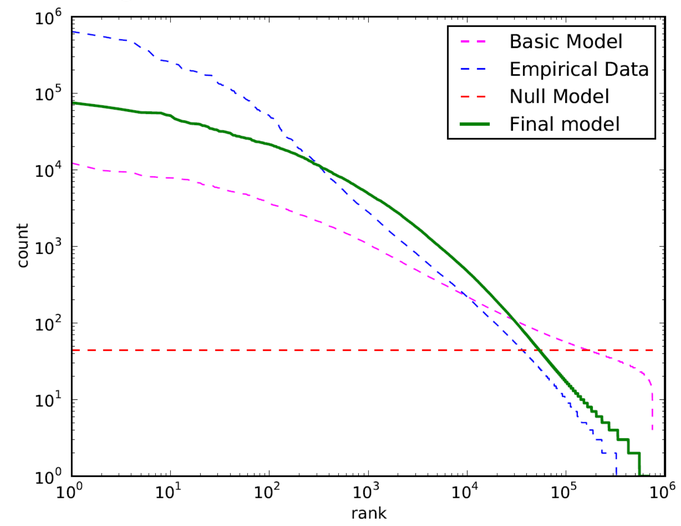
Frequency-rank plots of overall tag use for each of the three basic models, across a sampling of *P* values. The bold blue line shows the empirical tag distribution.

Refined model

We refined the top-5 model by adding a **memory component** This substantially improved model fit.

Model summary:

- With probability *P*, agents copy one of the top five existing tags for a particular item
- With probability *1-P* agents engage in novel tagging behavior:
  - With probability *Q*, assign a previously encountered tag with probability proportional to number of times encountered
  - With probability *1-Q* assign a random tag



Frequency-rank plots of overall tag use for the fitted version of the final, two-parameter model (P = 0.6, Q = 0.9), the original top-5 model (P = 0.6), the null model, and the empirical data

Conclusions

- Much social tagging activity may be driven by heuristic decision-making that manifests as **replication of existing popular tags**.
- These results should arouse some **skepticism around folksonomy practices**: They raise the possibility that the emergent terminology that develops within such systems may be partly driven by simple copying mechanisms.
- Future work will extend the model to explore if we can better predict tag popularity, as well as other metrics not explored here (e.g. tag co-occurrence)

Selected references

[1] Golder, S. A., & Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. Journal of Information Science, 32(2), 198–208.
[2] Schifanella, R., Barrat, A., Cattuto, C., Markines, B., & Menczer, F. (2010). Folks in folksonomies: social link prediction from shared metadata. WSDM, 271–280.
[3] Cattuto, C., Loreto, V., and Pietronero, L. Semiotic dynamics and collaborative tagging. PNAS 104, 5 (2007), 1461–1464.
[4] Vander Wal, T. Explaining and Showing Broad and Narrow Folksonomies, 2012.
[5] Todd, P. M., and Gigerenzer, G. Ecological Rationality: Intelligence in the World. Oxford Univ Pr, 2011.