
Answering questions about tag spaces using graphs

What can we really conclude from seeing a long tail, power law, or more complicated curve?

Questions

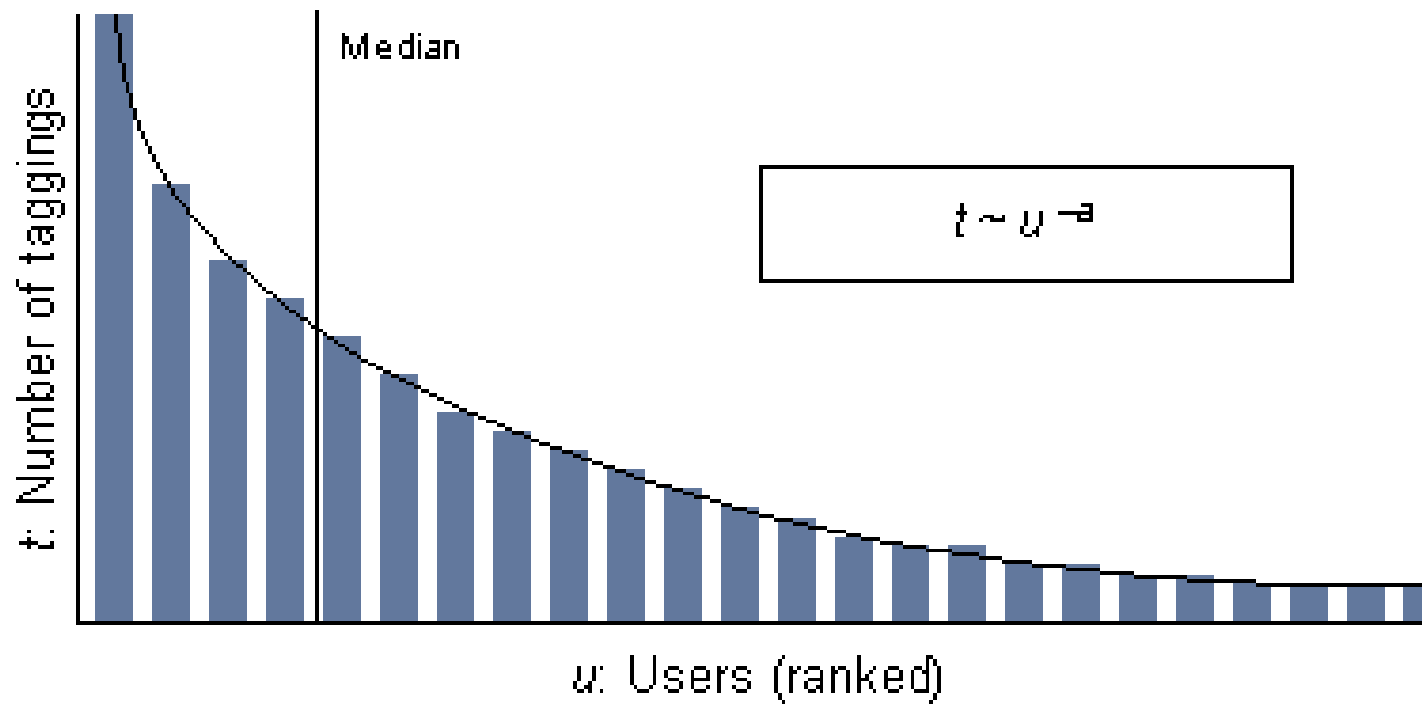
Global questions

- Are most taggings by the most active users?
- Are most taggings of the most popular URLs?
- Do most taggings use the most popular tags?
- Is there a meaningful average number of taggings per user (or per URL, etc.)?

Local questions

- Does a specific user mostly use a few top tags?
- Is a specific URL mostly described by a few top tags?
- Is a specific tag mostly used by a few top users?

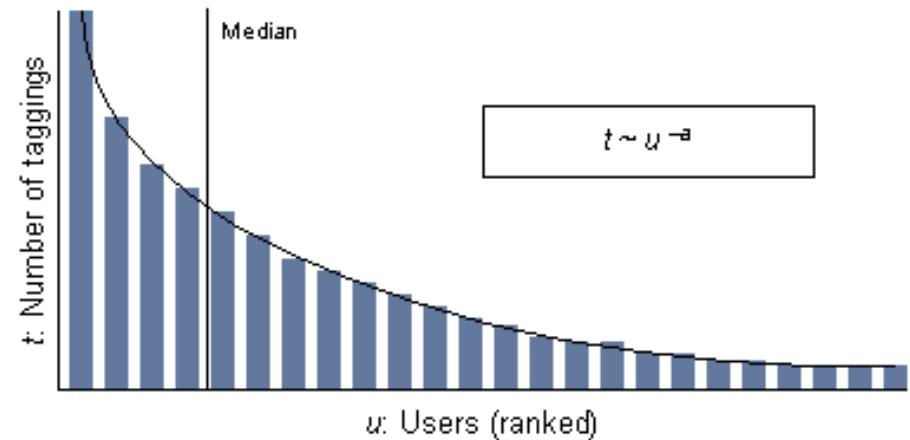
Answering with the “long tail” graph



Ranked histograms

Features

- Always decreasing
- Never a bell curve
- “Area” is total taggings
- Often fits a power law



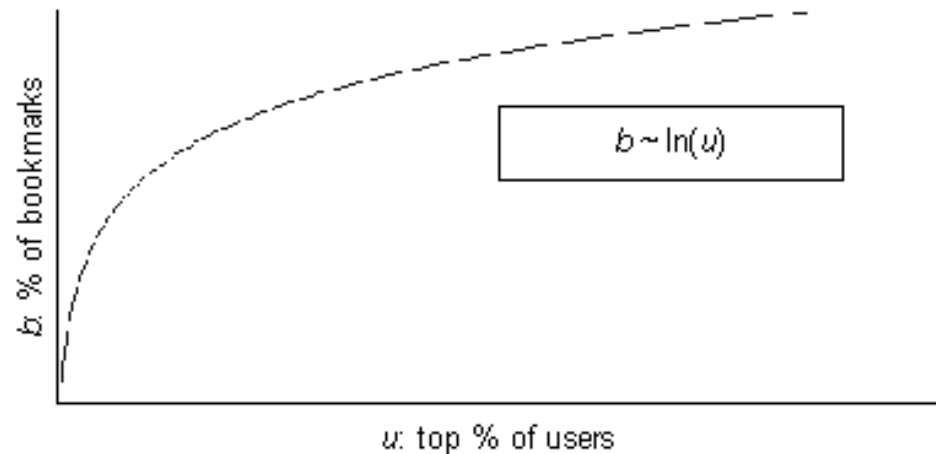
Answers

- Area to left / right of median is 50% of total taggings
- If only a few top users are left of the median, “most” taggings are made by the top users

Integrating for a more direct answer

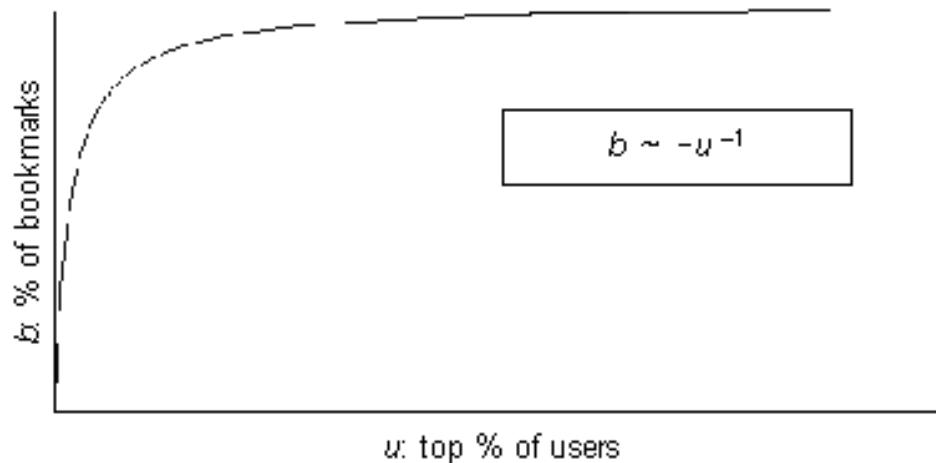
Integrating $t \sim 1/u$

- AKA “Zipf’s law”
- Gives a log

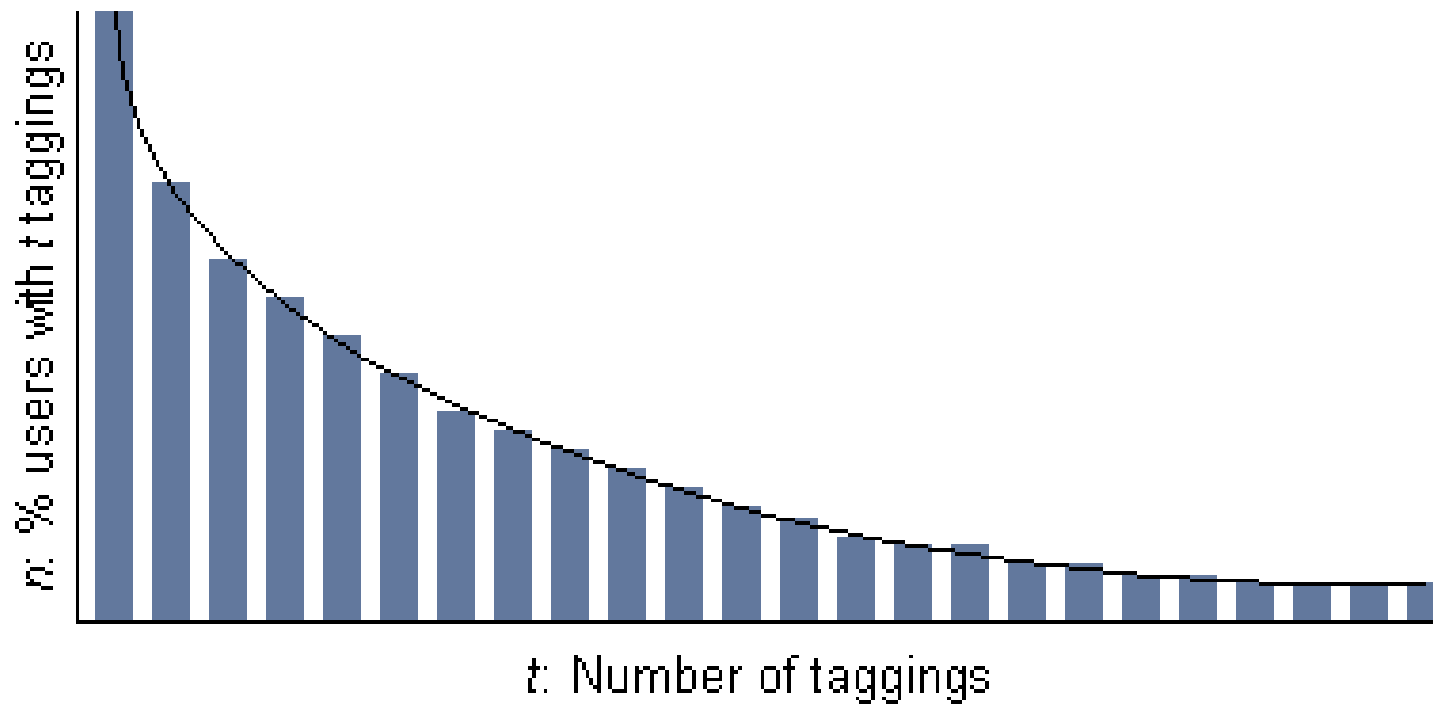


Integrating $t \sim 1/u^2$

- Gives a negative power law



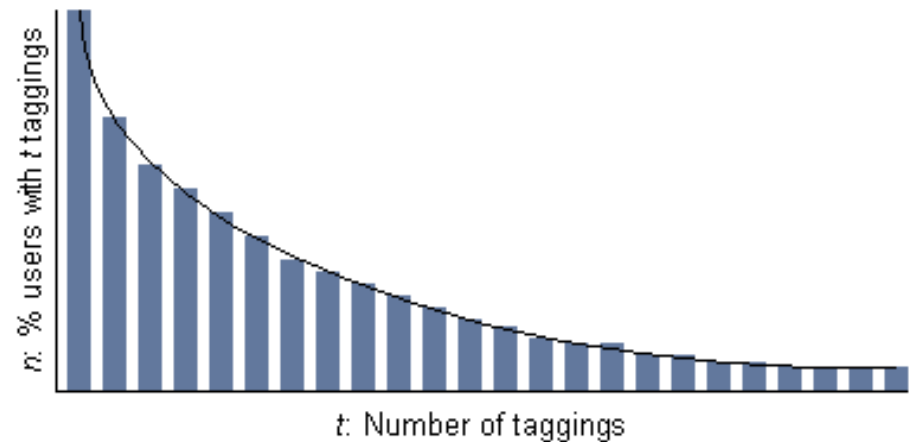
Answering with a probability distribution



Probability distributions

Features

- AKA PDF
- Can be any shape
- Directly shows most common numbers of taggings



Answers

- Average is “most meaningful” if it fits a bell curve
- If it fits a power law, average is relatively meaningless

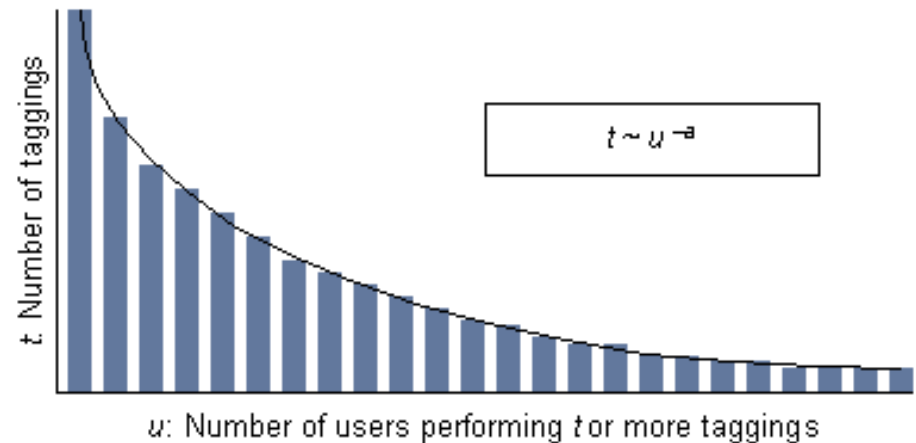
From ranking to PDF

① Equivalency:

“the u^{th} user did t taggings”

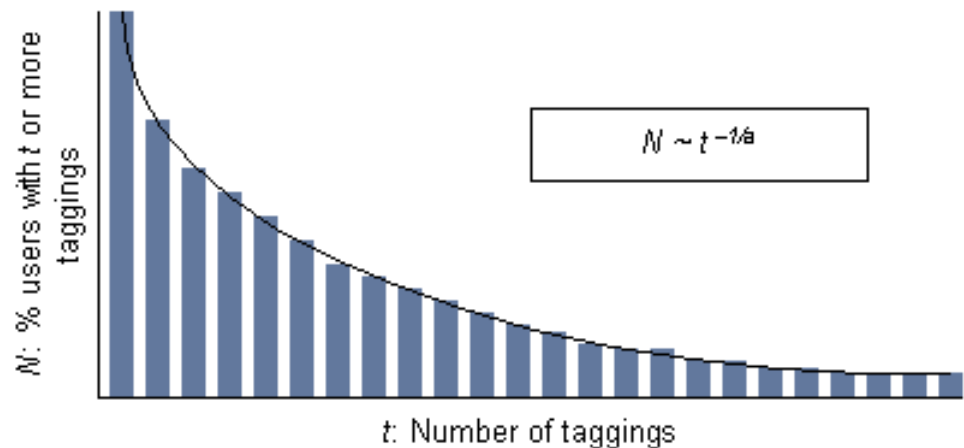
=

“ u users did $\geq t$ taggings”



② Invert

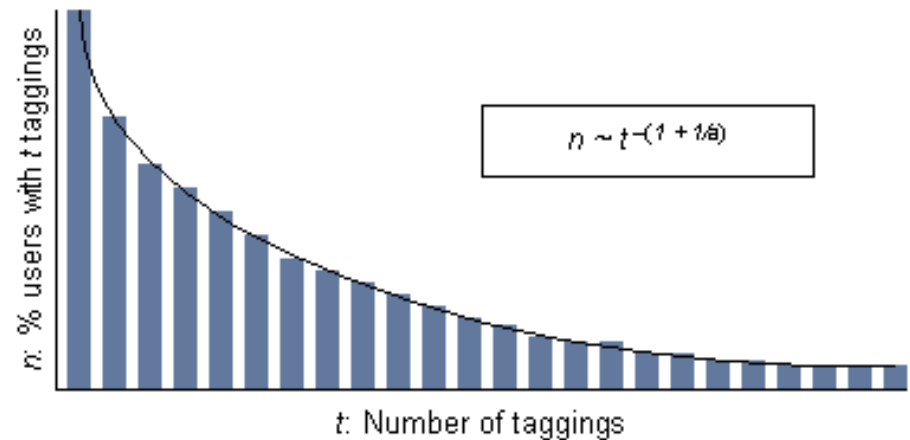
- This is a Cumulative Distribution Function
- AKA CDF
- AKA “Pareto” if power law



From ranking to PDF

③ Negative derivative

- # users who did $\geq t$ taggings minus # who did $\geq t + 1$ is # users who did exactly t



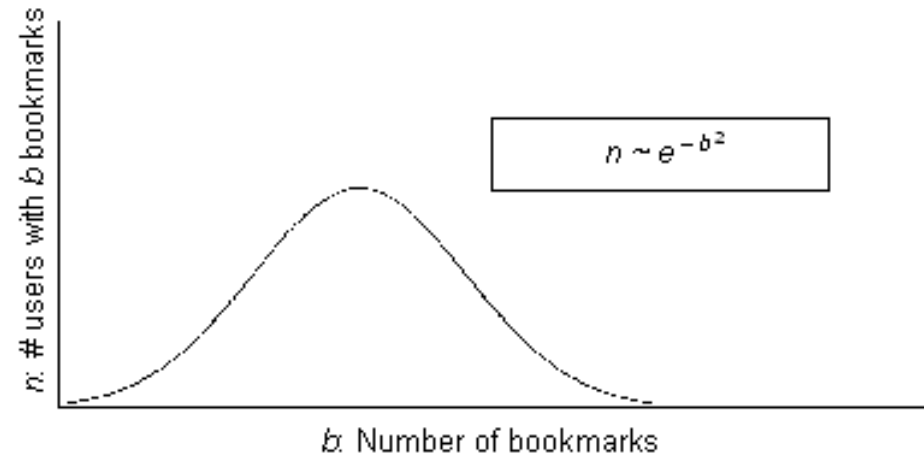
Answers

- If the ranked graph is a power law, so is the PDF
- So for a power law ranked graph, avg is meaningless
- Opposite is not true: PDF can be a power law, but ranked graph may not be

From bell curve PDF to ranking

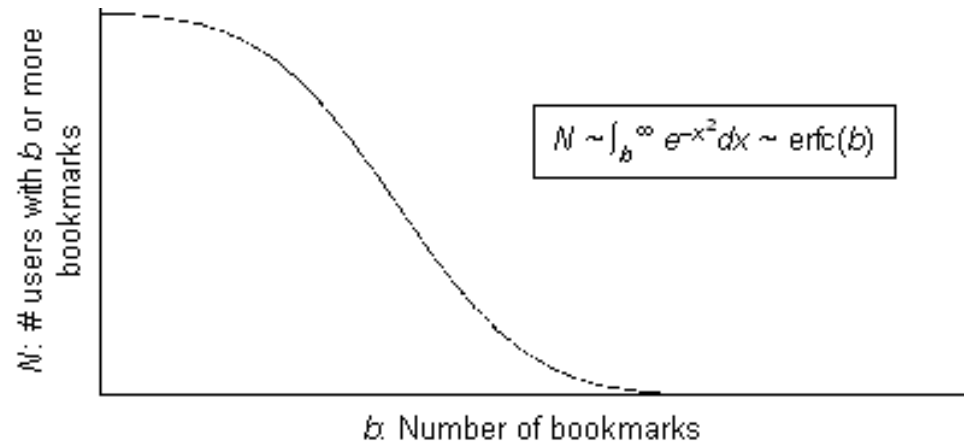
Features

- AKA Gaussian, normal
- Avg most meaningful



① Integrate

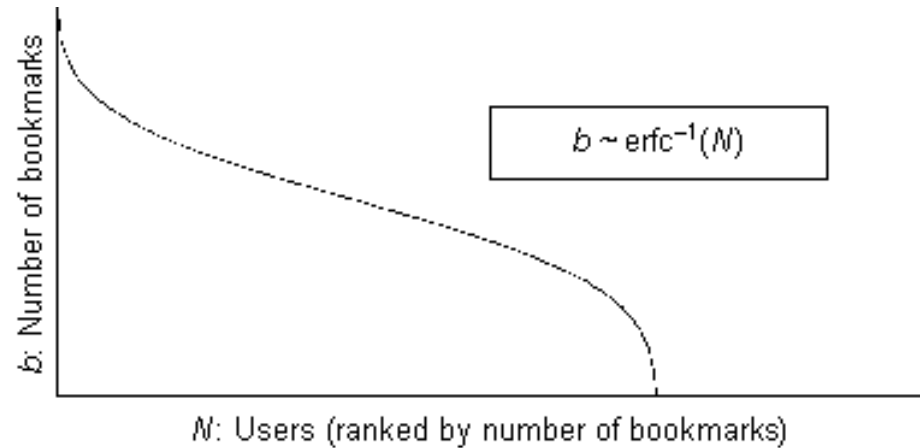
- Gives CDF
- Erfc is the “complementary error function”



From bell curve PDF to ranking

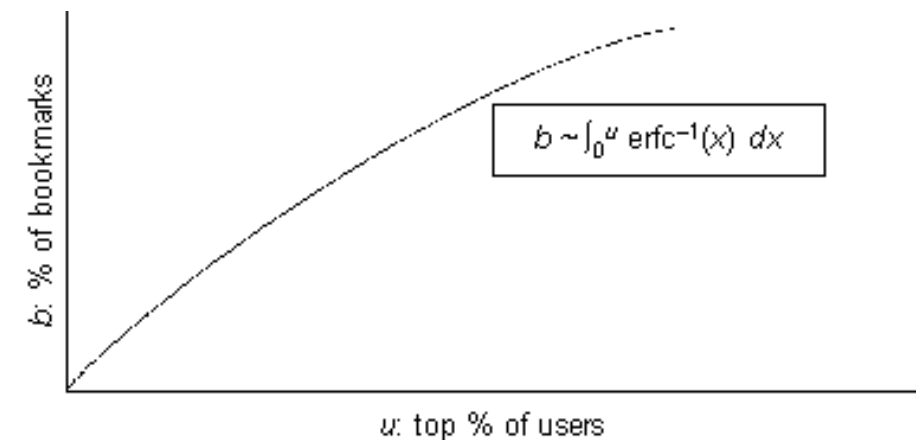
② Invert

- Exhibits a “long tail”, but is not a power law
- Has a maximally meaningful average

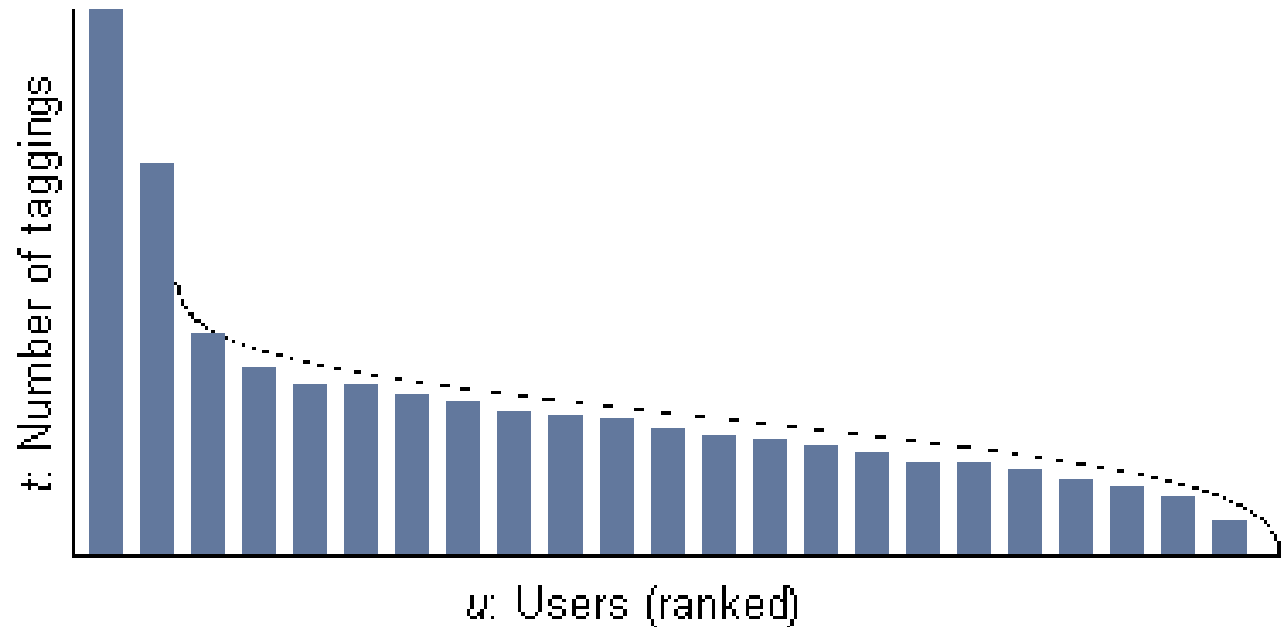


③ Integrate

- Top users do not have much of an outsized influence



Conclusion



Think twice about what that graph is telling you

- Long tail \neq power law
- Question being asked \rightarrow appropriate histogram
- Many curves can be fit to the same data