Burstiness and Memory in GitHub Activity Data

Nicholas Hanoian

GitHub API

- Each time someone **pushes** a change to a public repository on GitHub, it is recorded as an event on GitHub's public API
- We can use the time data associated with these pushes to learn about the patterns in which people do work
- We looked at over 150k repositories

The Dataset

- For each repository we have a set of times denoting when someone pushed to the repository

  $t(1), t(2), t(3), \ldots, t(n)$

- We can subtract the times from the previous time to get the time between events

  $\tau = t(i+1) - t(i)$

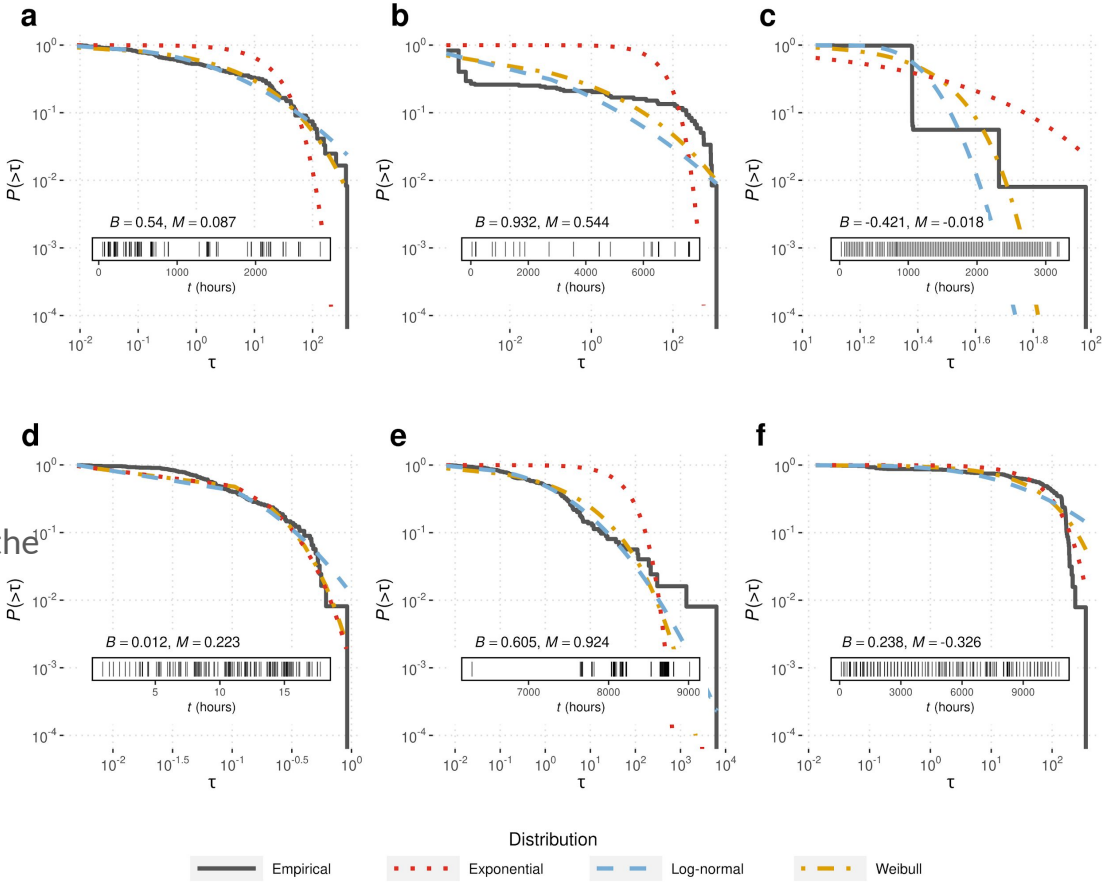
- We're interested in the distribution and associated measures of these interevent times

  $P(\tau)$

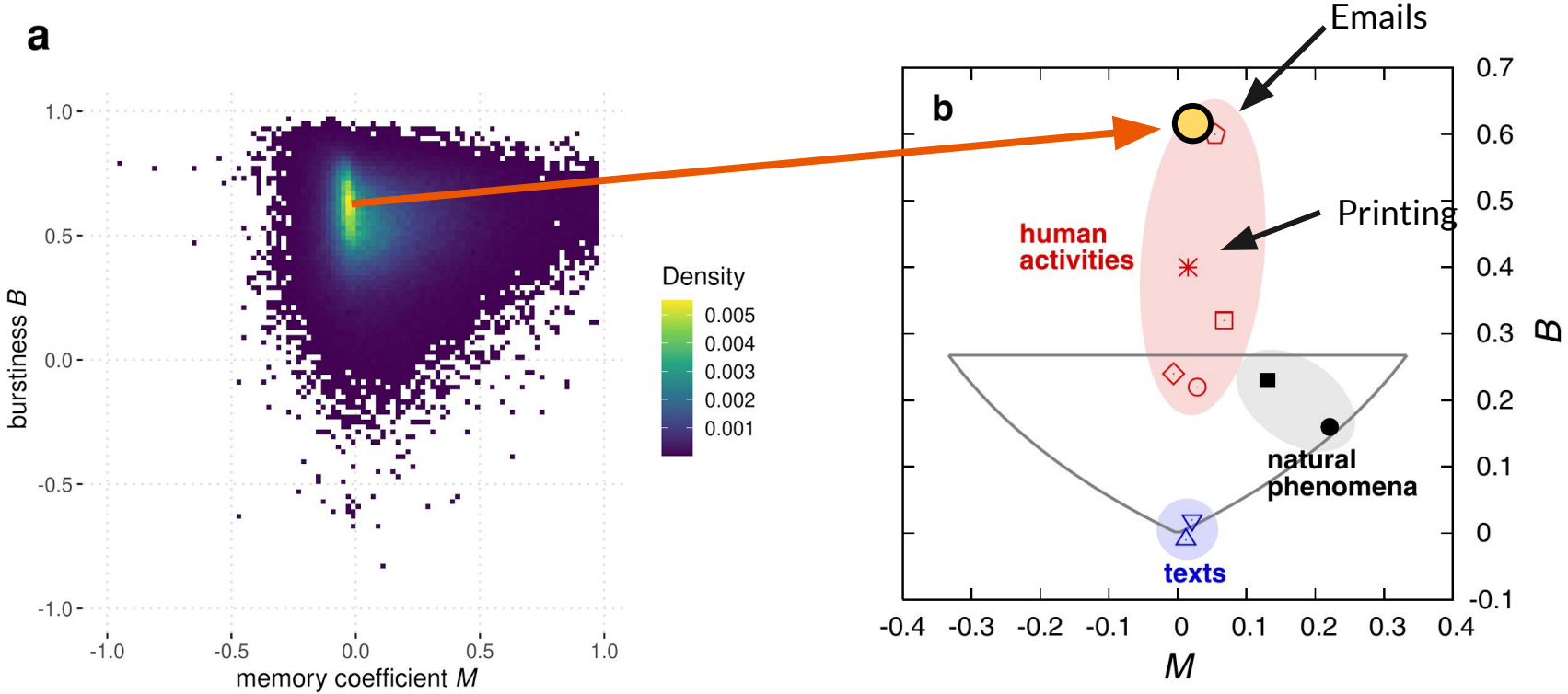
Burstiness and Memory

- Burstiness (*B*) is calculated using the mean and standard deviation of a probability distribution
- Ranges between -1 and 1
  - -1: anti-bursty (ex. heartbeat)
  - 0: random (Poisson process)
  - 1: bursty (human behavior)

- Memory coefficient (*M*) the autocorrelation of the signal
- Ranges between -1 and 1
  - -1: differing durations follow each other
  - 0: no correlation
  - 1: similar durations follow each other



Joint distribution of *B* and *M*



From Goh, Barabasi 2008