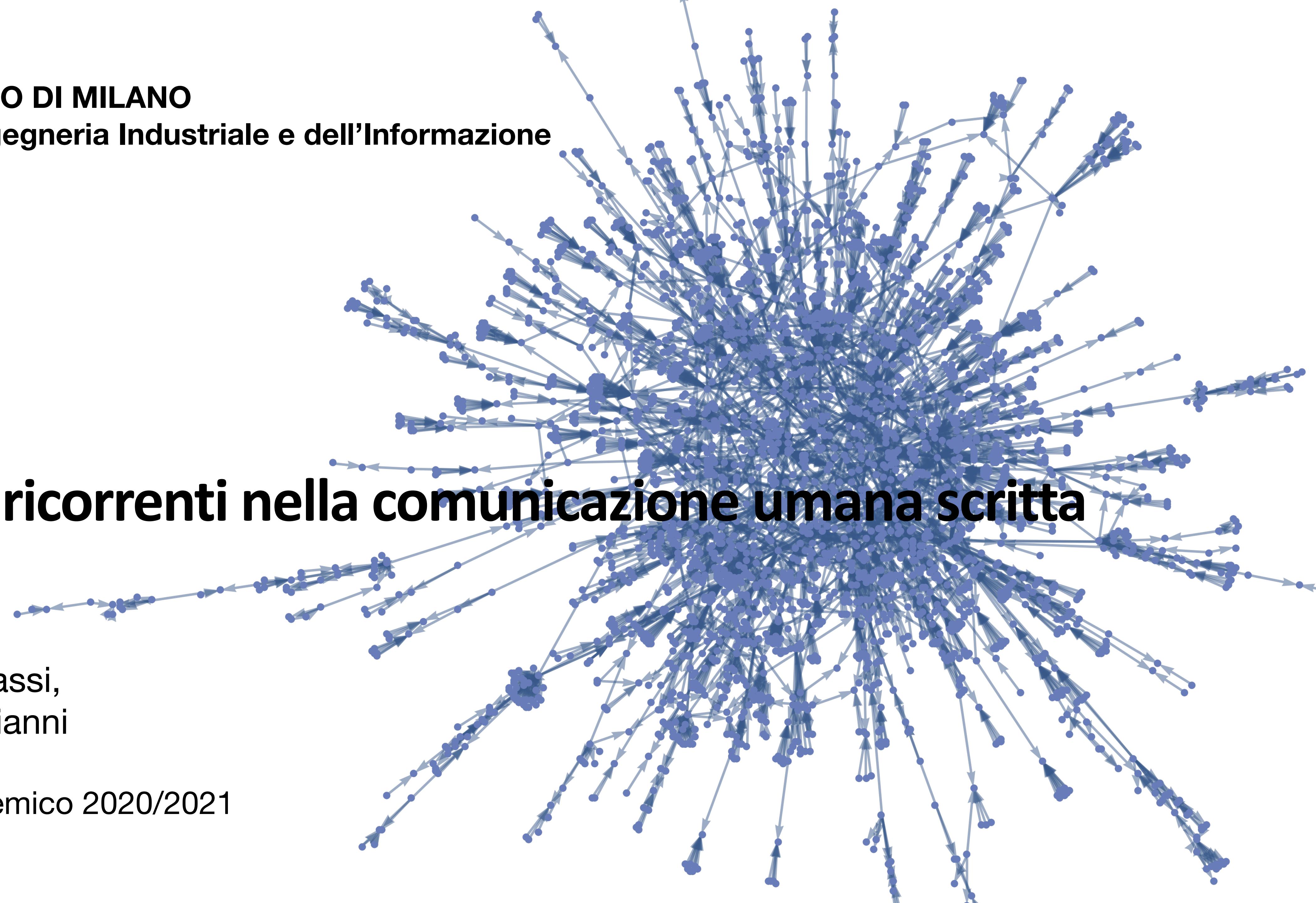




# **Pattern ricorrenti nella comunicazione umana scritta**

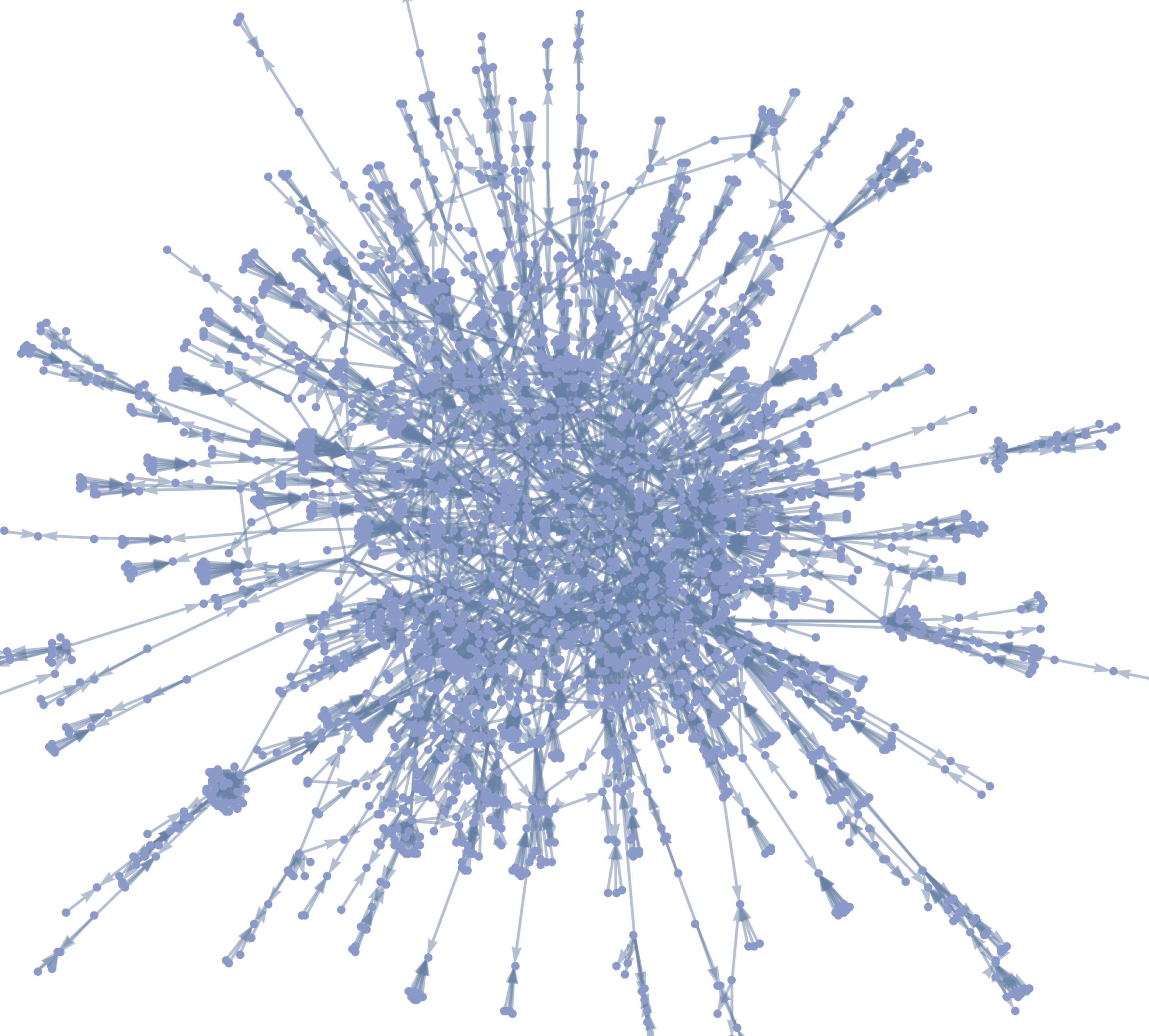
Giacomo Bassi,  
Olmo Notarianni

Anno Accademico 2020/2021



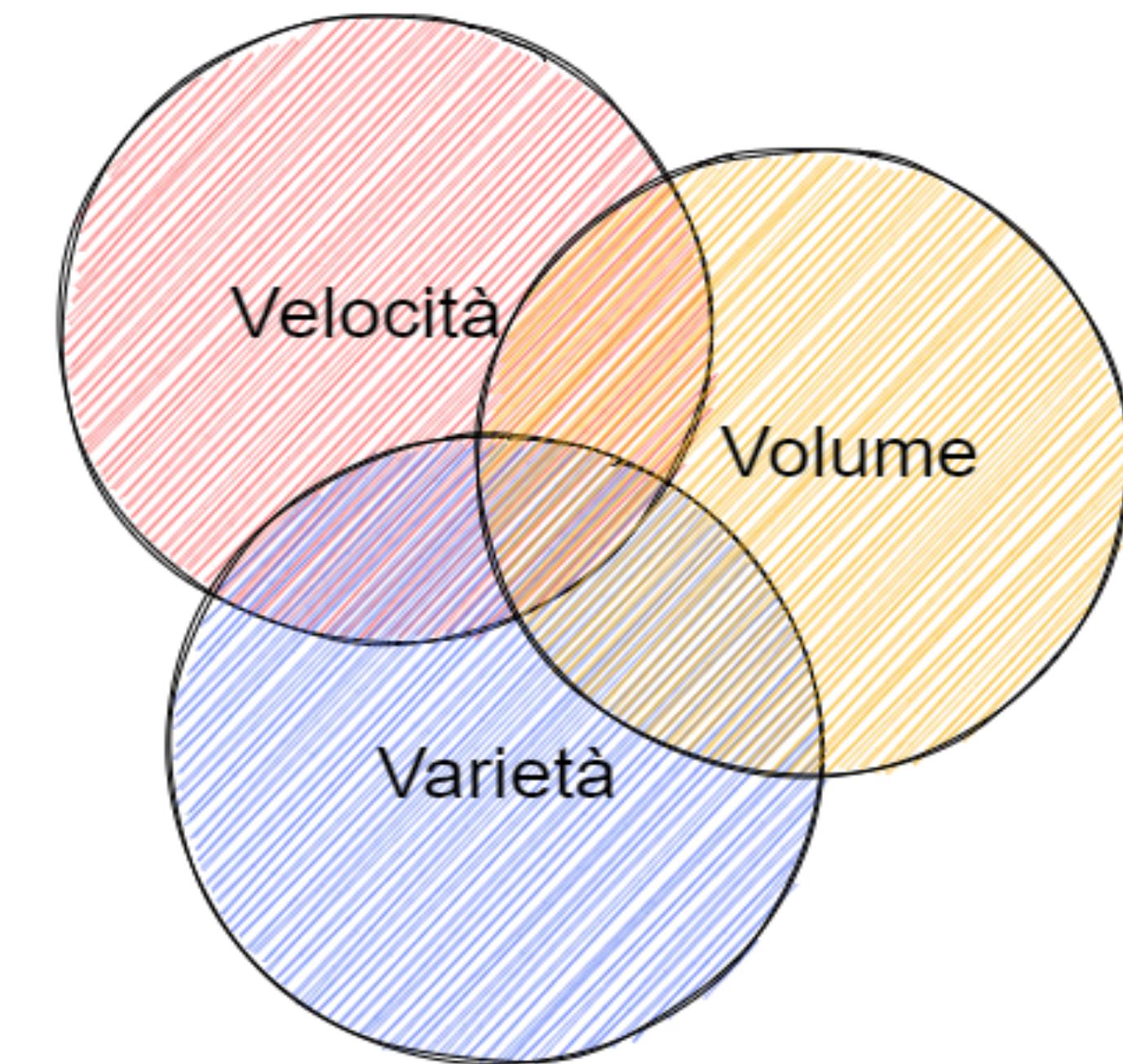
# Metodi e obiettivi

- Analisi di Big Data
- Wolfram Mathematica ®
- Protocollo Postfix
- Rappresentazioni
- Pulizia dei dati e Alias
- Selezione degli utenti
- Tempi e volumi di risposta
- Analisi comportamentale



# Il dato

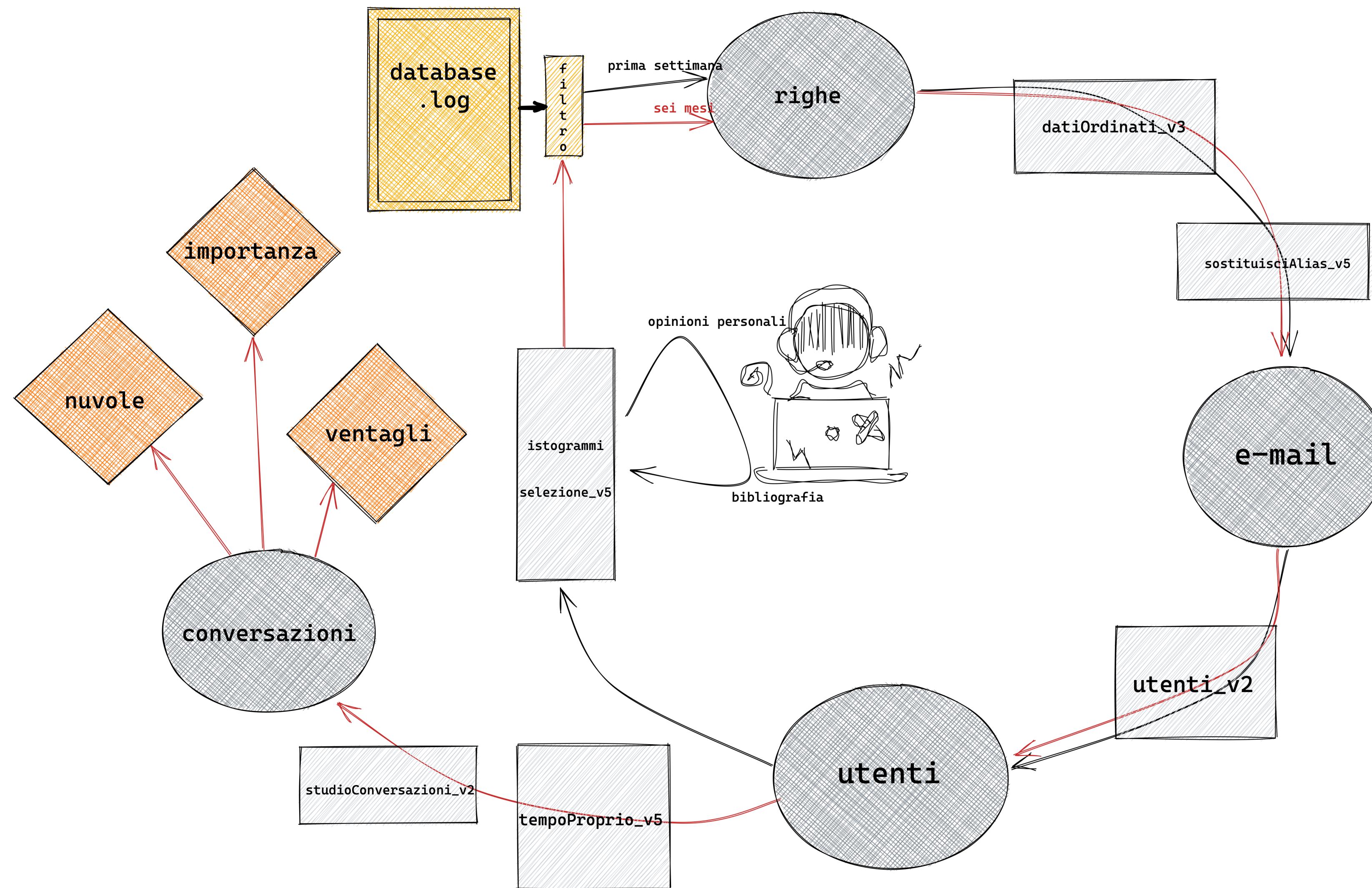
Oltre 50 milioni di righe registrate da tre mail server, a centinaia al minuto, per un anno.



In figura le “tre V dei Big Data” secondo Doug Laney e una porzione del nostro dato di partenza

```
Oct  9 06:29:49 bronze postfix/cleanup[2573]: 3E3391119F: message-id=<600fa399bb503e2f6a6c3629c7283623>
Oct  9 06:29:49 bronze postfix/qmgr[13771]: 3E3391119F: from=<cf949042d0c8e35657c0000d2c884957>, size=13026,
Oct  9 06:29:49 bronze postfix/lmtp[2392]: 3E3391119F: to=<45dbe63a8cbd60f8d790beff8f071d93>, relay=85c1228
Oct  9 06:29:49 bronze postfix/smtp[32386]: 3E3391119F: to=<57c68bf8545a4428cad4e86401d0d9d1>, orig_to=<45db
Oct  9 06:29:51 bronze postfix/cleanup[2573]: 16F0C1119F: message-id=<98f9e0adc4ff381270a22dde356ac294>
Oct  9 06:29:51 bronze postfix/qmgr[13771]: 16F0C1119F: from=<086e6127e55bcf49cdb76abcb192540>, size=11365,
Oct  9 06:29:51 bronze postfix/lmtp[2574]: 16F0C1119F: to=<b5d140cb5754edbea6ec25c904a59e73>, relay=cd0b7fc
Oct  9 06:29:52 bronze postfix/cleanup[2573]: 1BACE1119F: message-id=<e083bf73b462a84b2a85b90853d576f3>
Oct  9 06:29:52 bronze postfix/qmgr[13771]: 1BACE1119F: from=<ed9261272d8bb318984baa32b7aadd1b>, size=12239,
Oct  9 06:29:52 bronze postfix/lmtp[2178]: 1BACE1119F: to=<a7a0a39c72c10e83f61b45f9ada84fa5>, relay=cd0b7fc
Oct  9 06:29:53 bronze postfix/cleanup[2573]: 0AE0D1119F: message-id=<333549f733edfbdd54d42527b30c4924>
Oct  9 06:29:53 bronze postfix/qmgr[13771]: 0AE0D1119F: from=<54a9e6e20948ebe96d80da6c139a37d0>, size=16881,
Oct  9 06:29:53 bronze postfix/lmtp[2574]: 0AE0D1119F: to=<878bf20b526642535835a6a88fa208e6>, relay=cd0b7fc
Oct  9 06:30:15 bronze postfix/master[11580]: reload -- version 2.9.6, configuration /etc/postfix|
Oct  9 06:30:15 bronze postfix/master[11580]: warning: ignoring inet protocols parameter value change
```

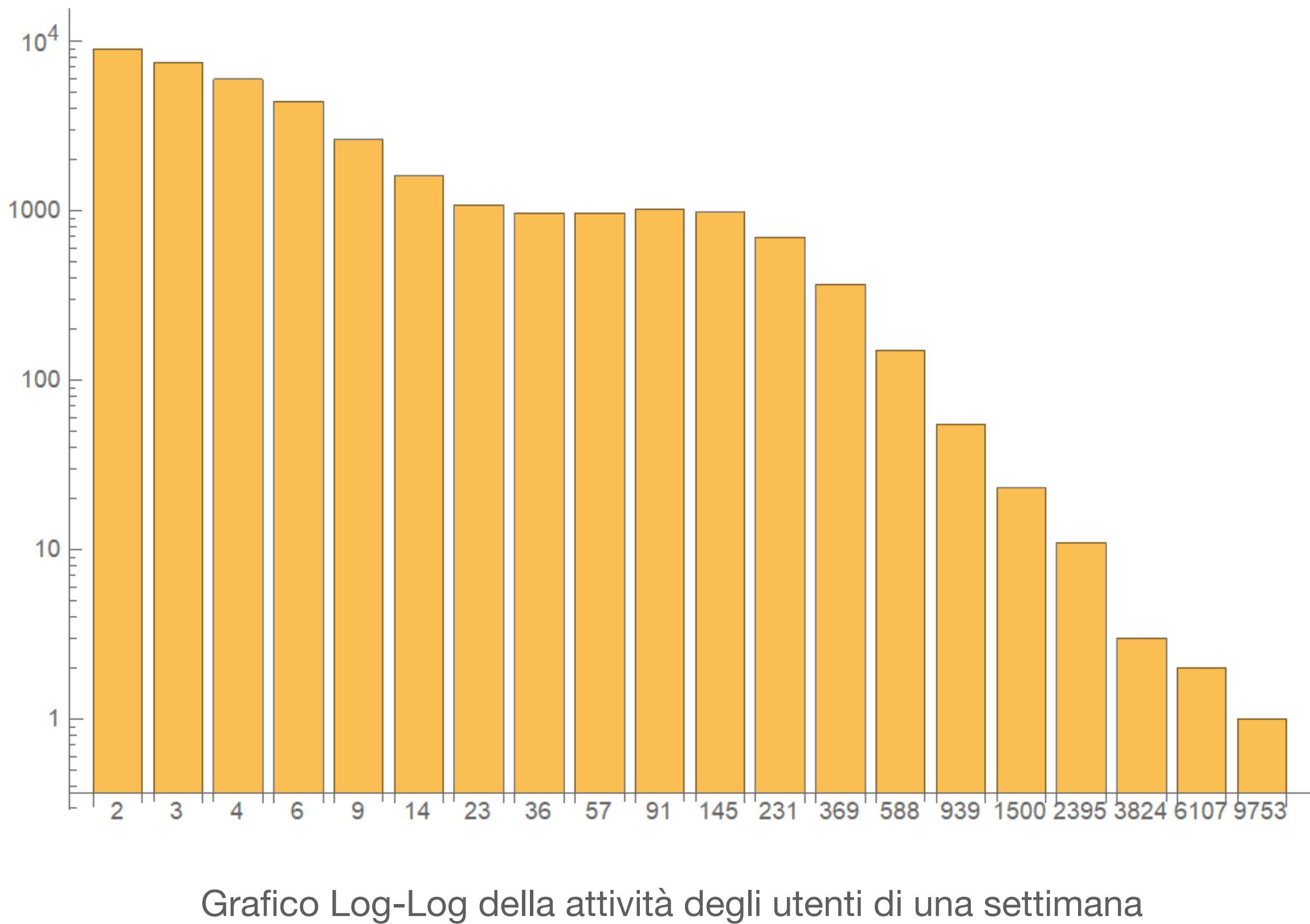
# Diagramma di flusso



- Linea nera:  
37200 utenti,  
una settimana;
- Linea rossa:  
10 utenti,  
sei mesi.



# Selezione degli utenti: il parametro Attività



L'attività è il principale quantificatore del comportamento degli utenti. Oltre a questa, studiamo il numero degli alias, il rapporto  $\rho$  (inviate/ricevute) e la dimensione media in B dei messaggi inviati.

$$\hat{(attivita)} = \frac{2}{3}(inviatepesate) + \frac{1}{3}(ricevutepesate), \text{ con}$$

$$(pesoemailinviata) = 1 + \text{Log}(\text{numerodidestinatarieeffettivi})$$

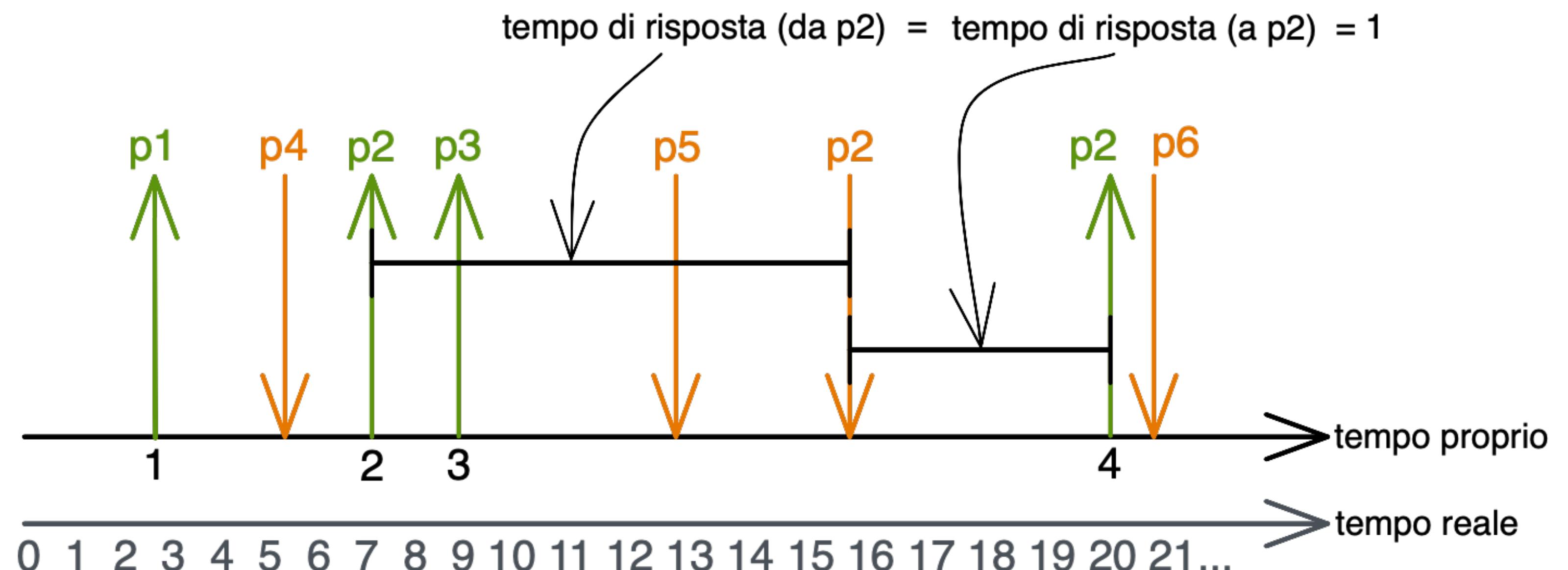
$$(pesoemailricevuta) = \frac{1 + \text{Log}(\text{numerodidestinatarieeffettivi})}{\text{numerodidestinatarieeffettivi}}$$

# Tempo proprio

Un re-clocking scandito dall'attività dell'utente è più indicativo del tempo reale.

Le pause periodiche e “fisiologiche” vengono trascurate.

Il tempo proprio cresce di un'unità ogni volta che viene inviata una e-mail.



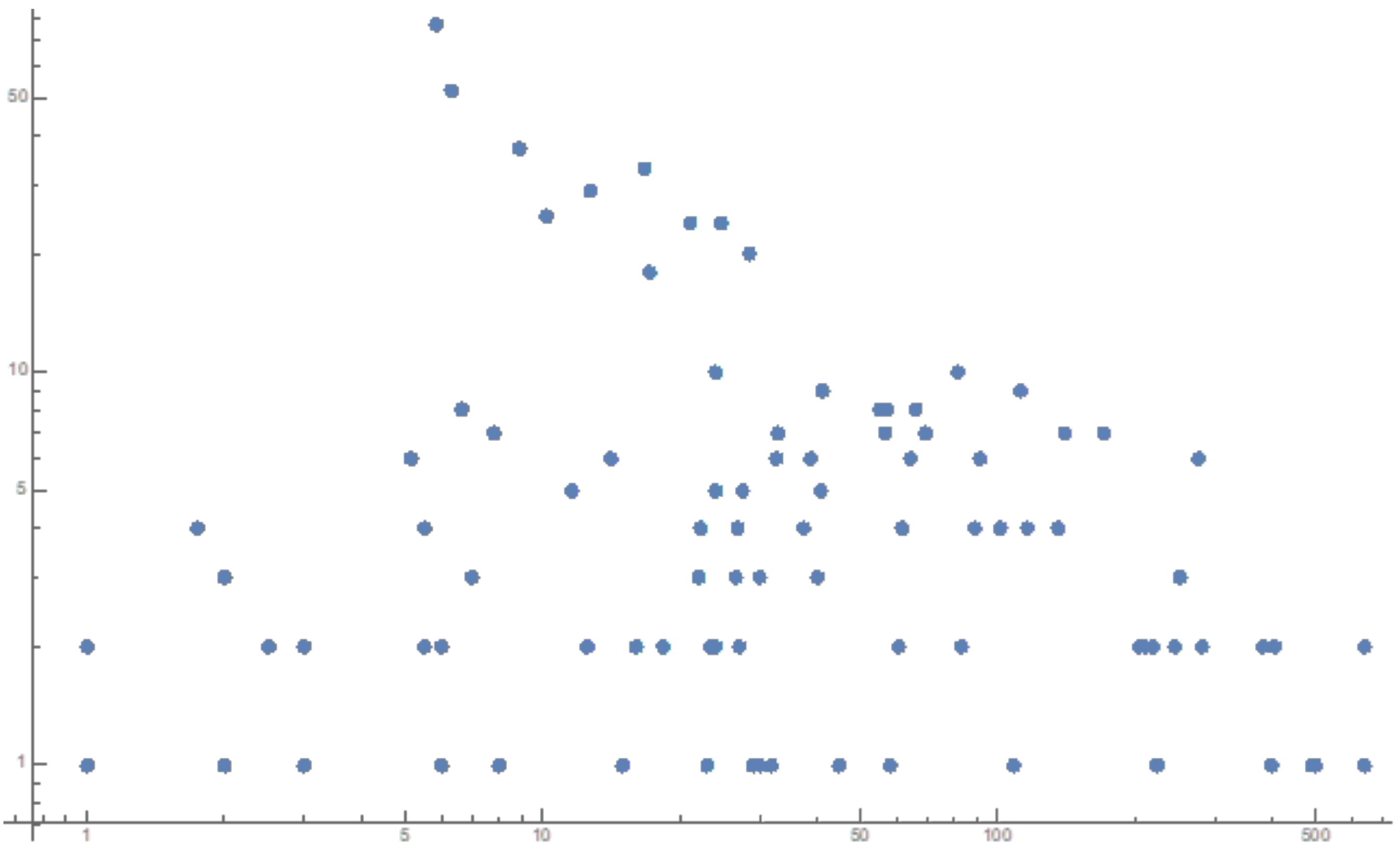
# **Correlazione tempi-volumi di risposta**

# Nuvole: intuizione visiva e coefficiente di correlazione di Spearman

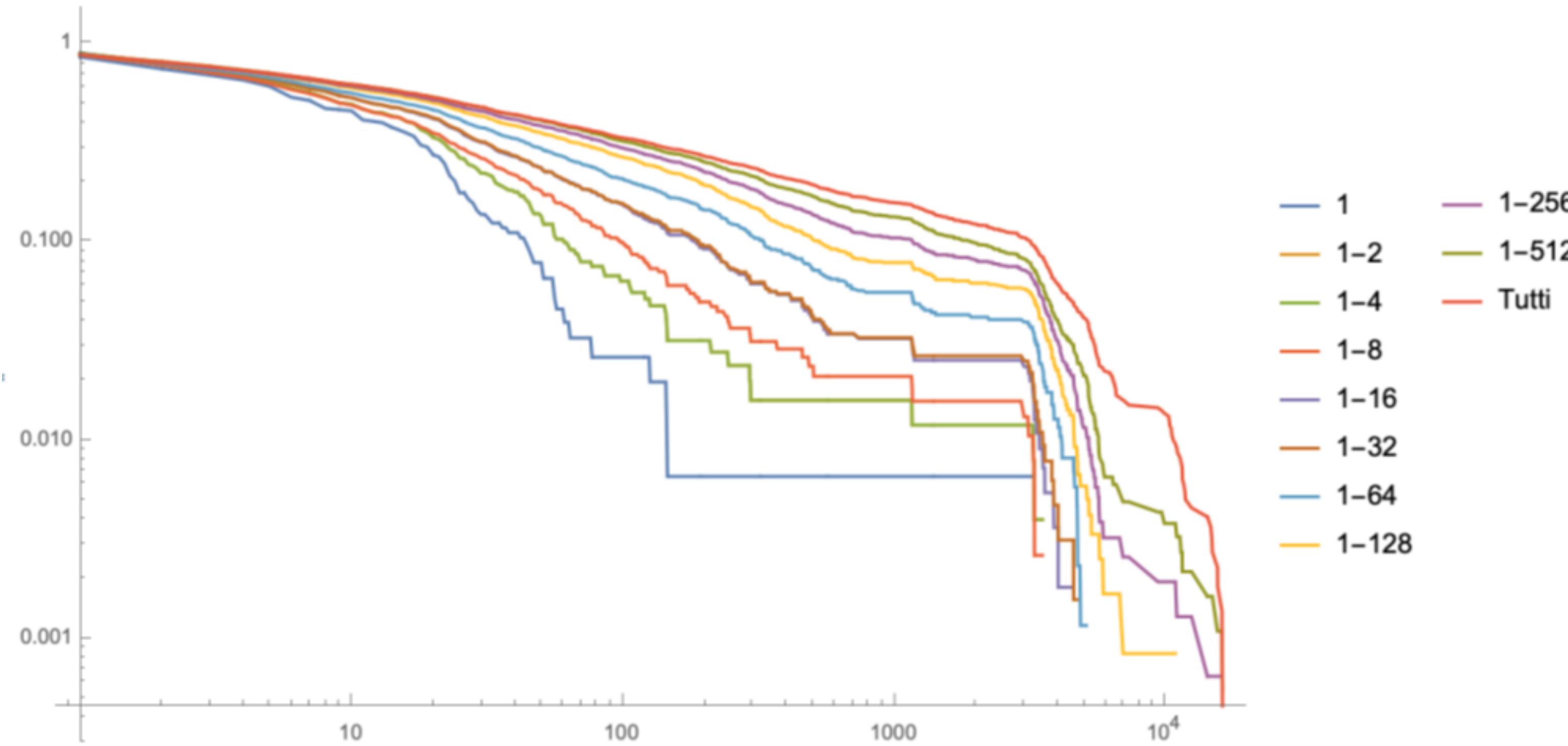
La distribuzione in scala logaritmica dei volumi scambiati rispetto ai tempi di risposta assume sempre la forma caratteristica che chiamiamo *nuvola*.

Ogni nuvola è costituita da due sottosinsiemi:

- Il bulk ( $y \leq 4$ ): partner con volumi piccoli e tempi uniformemente distribuiti
- Il picco ( $y \geq 4$ ): partner con volumi grandi e tempi di risposta bassi



# Ventagli: distribuzioni cumulate inverse dei tempi di risposta

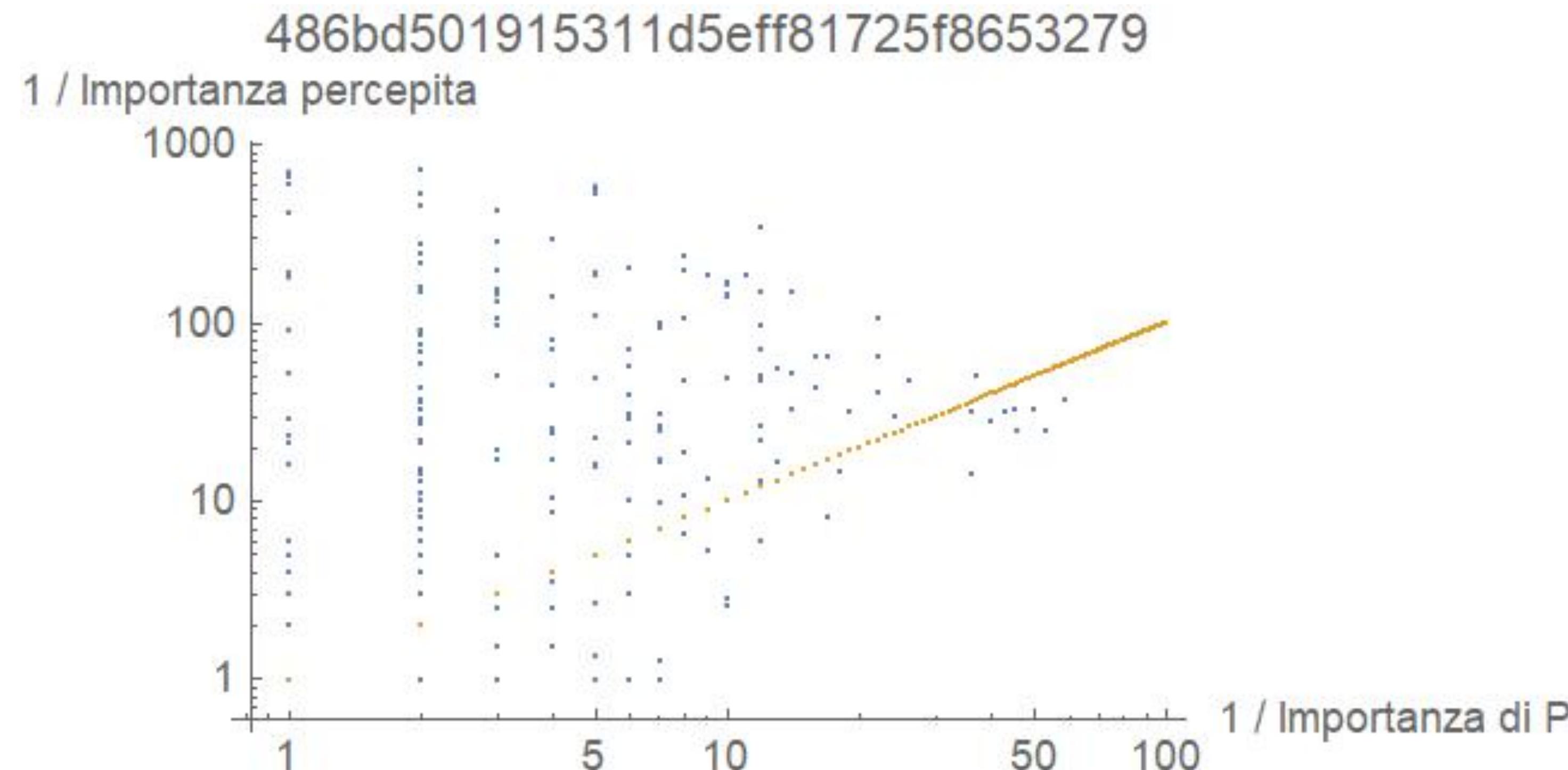


I grafici “a ventaglio” mostrano una correlazione decrescente tra volume scambiato e tempo di risposta medio.

I partner di un utente vengono ordinati dal più al meno contattato. La figura mostra l'aumento della probabilità di ricevere risposte in tempi lunghi considerando partner sempre meno contattati.

# Il rapporto utente-partner: Importanza e Importanza percepita

Tendiamo a dare importanza all'altr\* in base alla percezione della priorità che l\*i ci riserva?



A lato: la rappresentazione Log-Log dell'Importanza percepita rispetto all'Importanza data. Il riferimento giallo è la bisettrice del grafico.

(mediamente, l'utente «486b...79», percepisce l'altr\* più lent\* nel rispondere di quanto l\*i non sia).