



# **Scribee Experimentation**

## Early Statistics on Email Conversations

Vincent Verdot – Bell Labs

August 2011

# AGENDA

1. Context
2. The Scribee project
3. Experimentation details
4. Results
5. Conclusions

# CONTEXT

## Information overload and Email

- *Information overload*, or how too much information kills information, is today one of the major challenges the IT sphere has to address.
  - Leveraged by the Web 2.0's open model, everyone produce data,
  - Impact all communication mediums, Email is also concerned.
- As we cannot avoid this information overflow, how could we take advantage of this situation?
  - Filter, extract, generate knowledge from this information,
  - Adapt email client applications to help users acquiring this knowledge.

# CONTEXT

Some numbers (year 2010)

- **107** trillion emails sent  
= **294** billion messages per day
- Spam is estimated to represent **89%** of emails  
= **34** billion messages sent every single day

# THE SCRIBEE PROJECT

## Presentation

- *Scribee* is the name of an internal research project started in 2009 within the Bell Labs' Application Domain.
- It is a multidisciplinary project which brings together researchers, engineers, designers and sociologists.
- The objective is to study the email service behavior and their users' productivity in an information overload context.
  - Enabling the user to handle "One Million Conversations"
  - Building knowledge from email conversations

# THE SCRIBEE PROJECT

## An on-going study of Conversations

- Our research works are focused on the concept of email Conversation that we defined as a set of logically connected email messages
  - Messages are linked when sharing common characteristics: participants, discussion thread, subject, date and time, etc.
- This study did not provided yet its final conclusions but the experiments we conducted provided interesting values that should be taken into consideration when designing email-related applications.
- The current presentation will focus on early results corresponding to general email metrics that we collected during our experimentation.



# EXPERIMENTATION DETAILS

## Presentation

- We conducted an experiment within Bell Labs' Application, open to all researchers during 6 months
  - 158 involved, 65 active, all familiar with computers and email services.
- No installation was required and every user was free to use or not the prototype at anytime
  - By adding a specific email address (e.g. scribee@bell-labs.com) in the message's destination (To, Cc or Bcc fields).
  - Fully transparent for the email delivery.
- The prototype provided email-related services accessible through a dedicated web page, it will not be described here.

# EXPERIMENTATION DETAILS

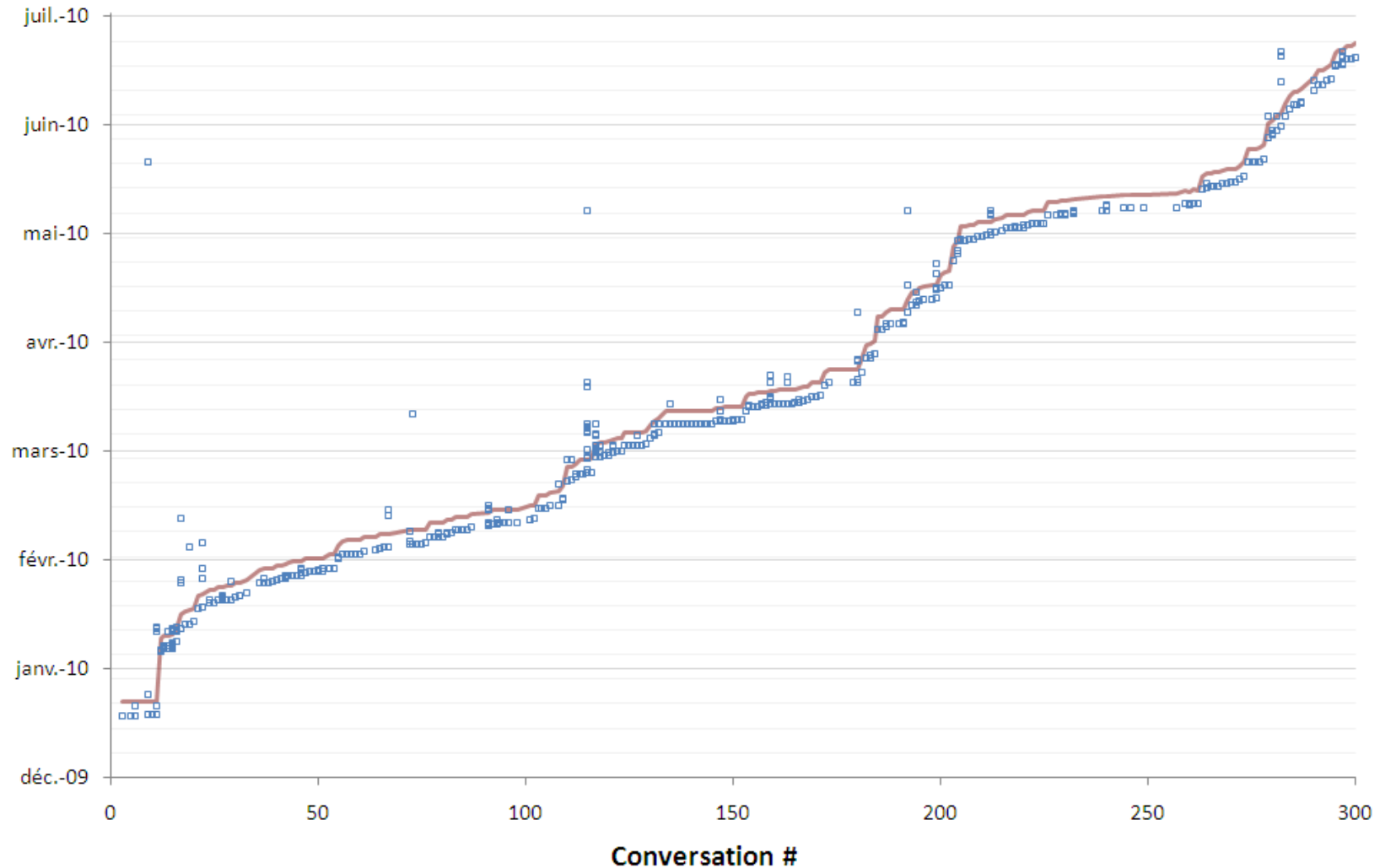
## Data gathering process

- Listening for every email message, the Scribee prototype collected various information about the conversations, processed online and offline.
  - That's the results presented here.
- Online by incrementing counters
- Offline by anonymously storing every received message (+ various metadata) for further processing
  - Messages are merged into conversations
- Some "cleanup" was performed to remove testing data
  - We kept about 600 messages / 300 conversations.



# RESULTS

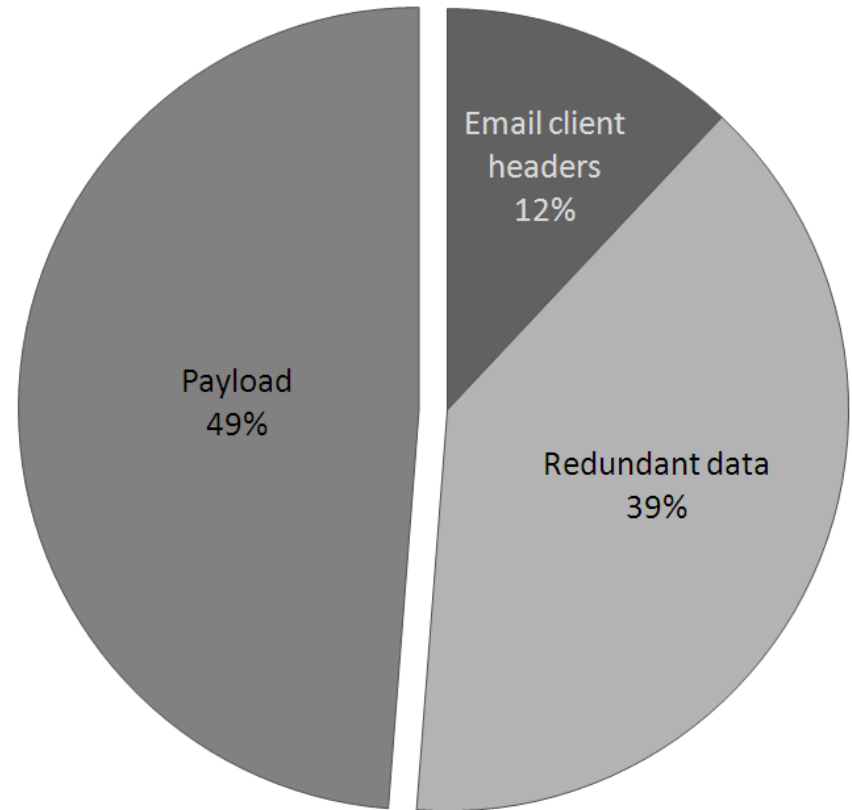
## Experiment activity timeline



# RESULTS

## Conversation payload

- We define “payload” as the useful part of a conversation, *i.e.* excluding redundancy
  - computable and cloned data,
  - metadata, headers and other system-related data are not redundancy
- The current implementation of email clients infers a lot of redundancy
  - separators, recursive copy of previous messages, etc

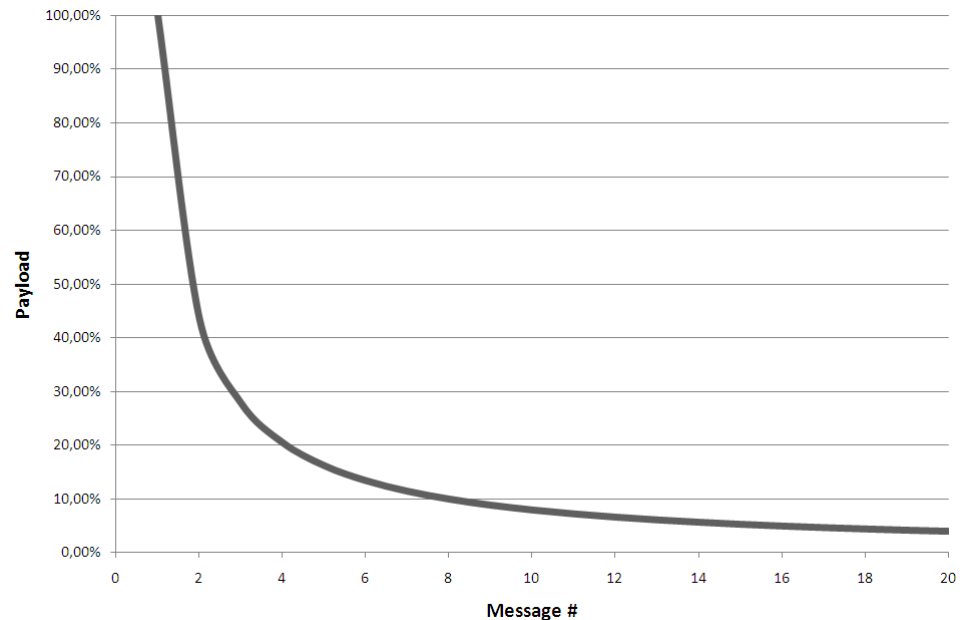


# RESULTS

## Conversation payload – Theoretical efficiency

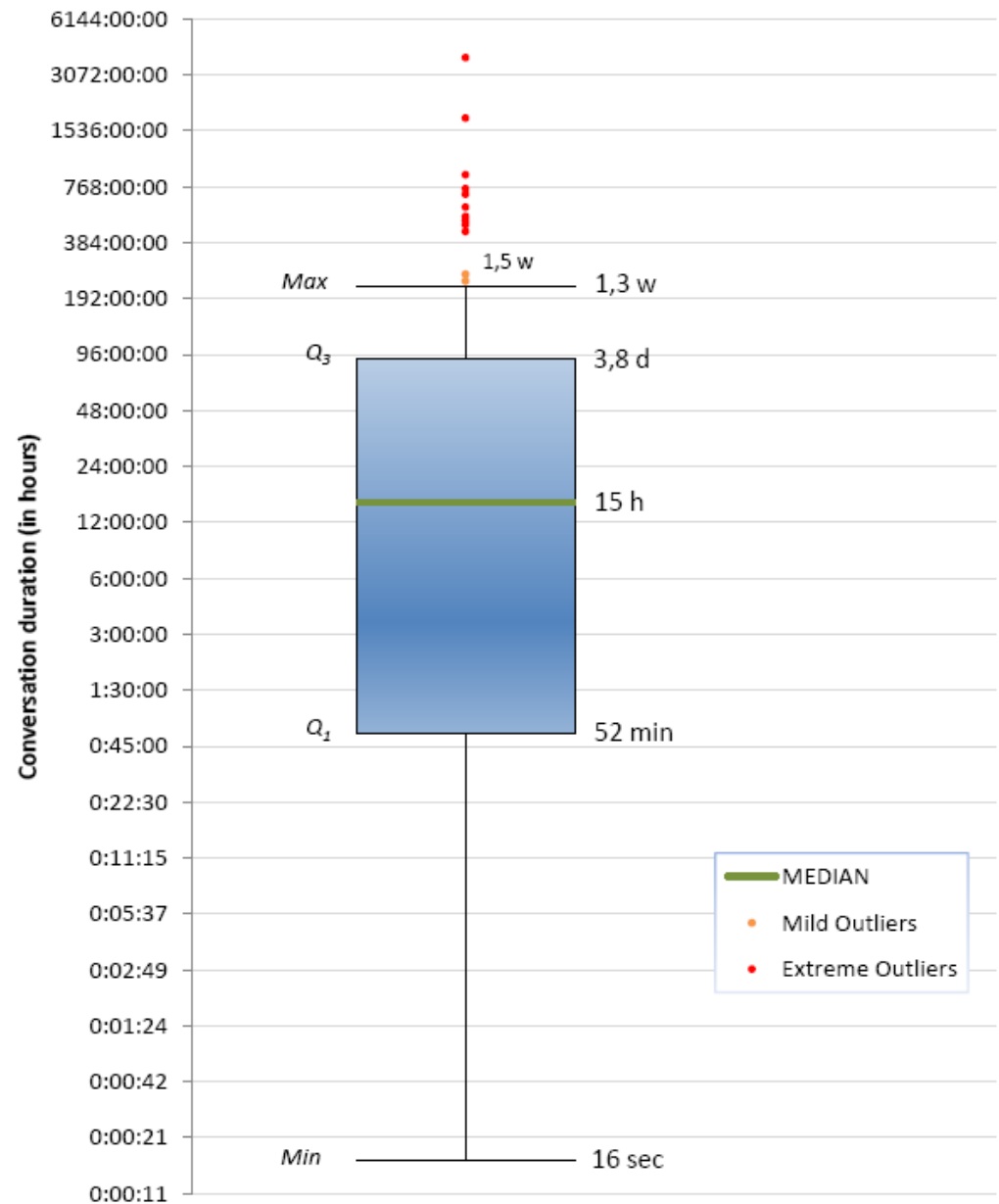
- The redundancy is cumulative for every new message added to the conversation
  - the payload of the nth message of a conversation can be estimated as follows (with  $S_n$  the size of the nth message and  $H_n$  its header size)
- Representation of the theoretical message efficiency within a conversation with average values we measured
  - The 5<sup>th</sup> message's payload only represents 15% of the message size

$$P_n = \frac{S_n}{\sum_{j=1}^n (S_j + H_j)}$$



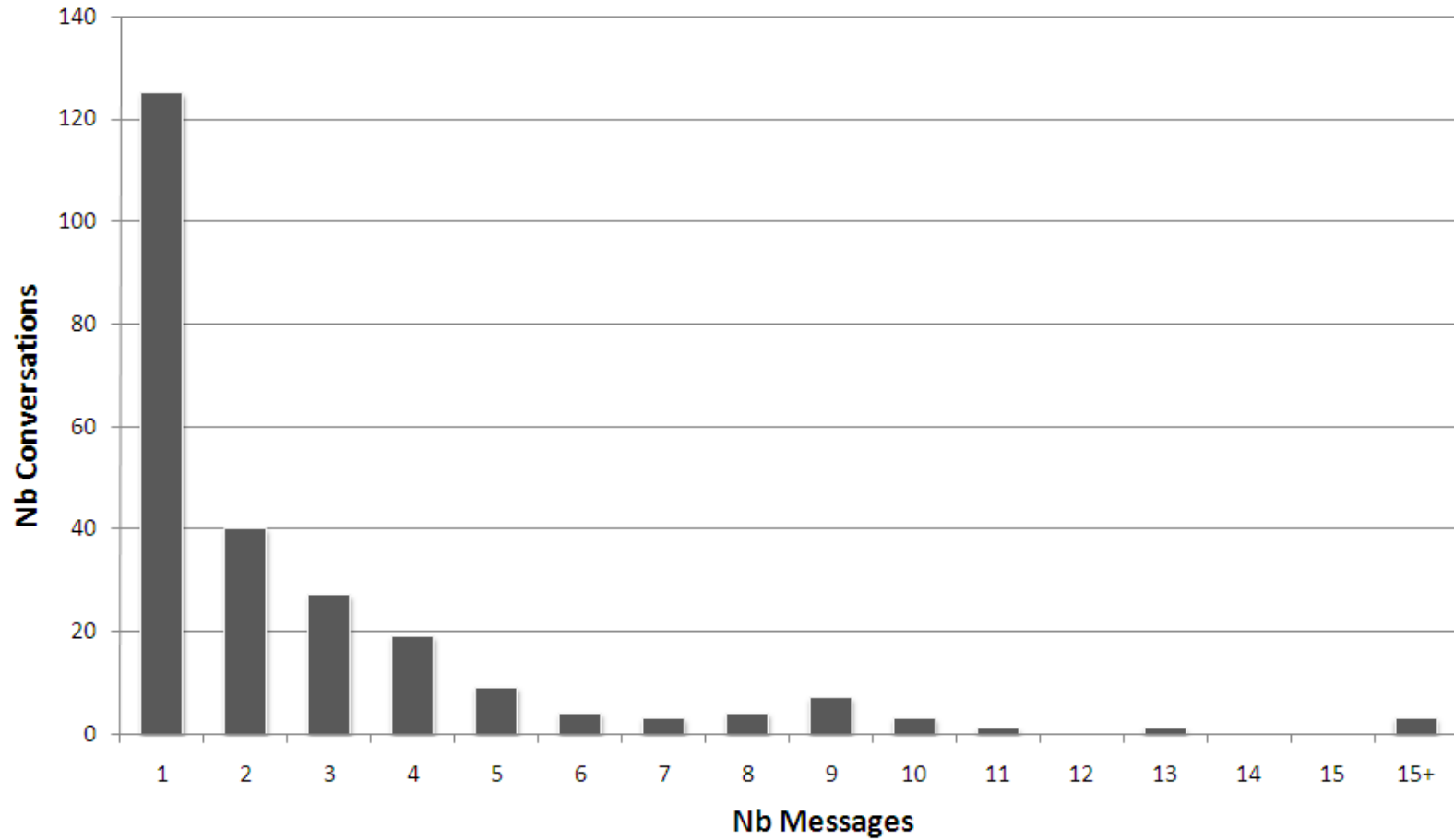
# RESULTS

## Conversation duration



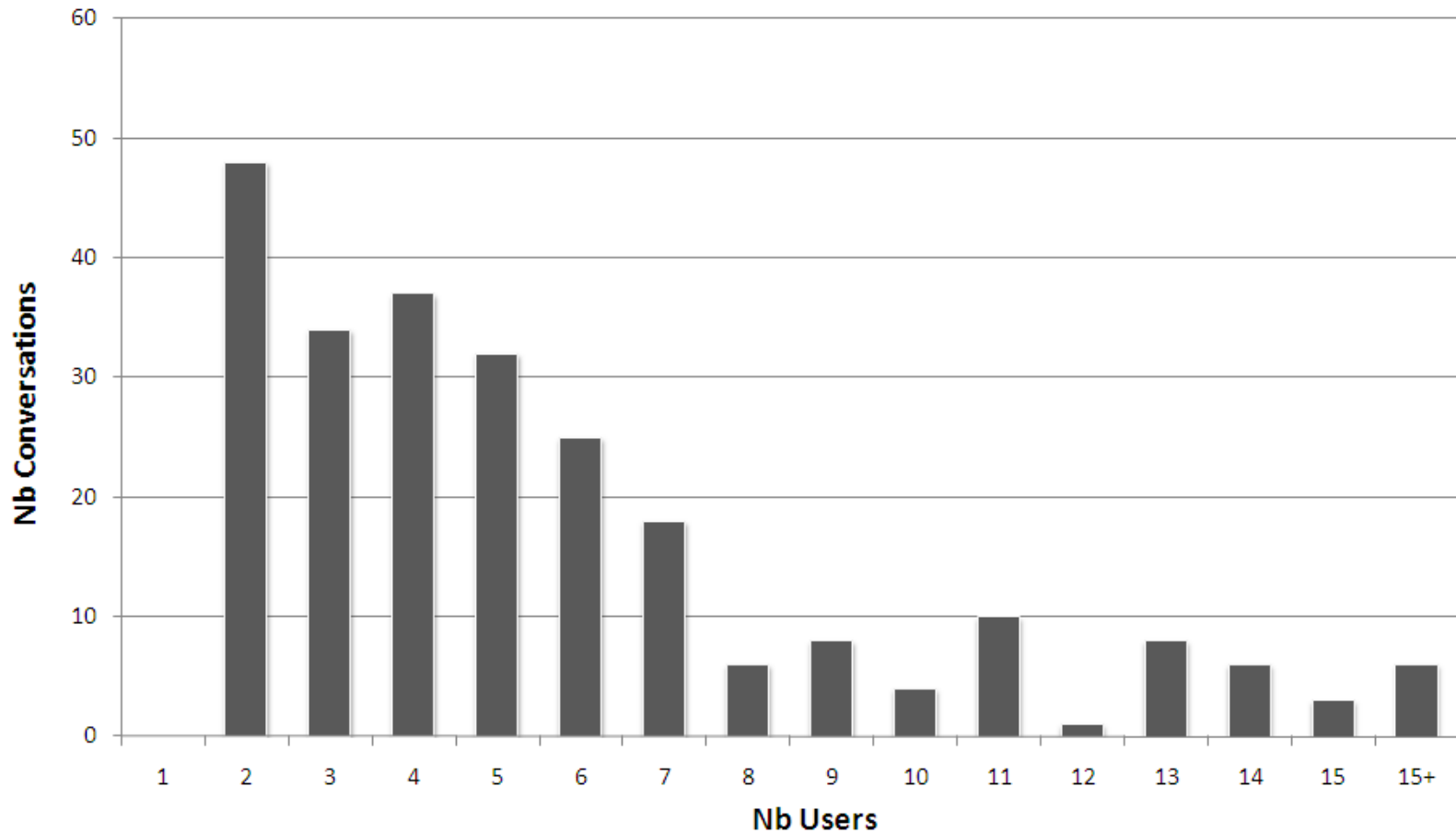
# RESULTS

## Additional measures



# RESULTS

## Additional measures



# CONCLUSIONS

## Conclusion and future works

- Only early results but yet interesting for email application design (and maybe all message-based communication services)
  - Revealed valuable properties of email conversations,
  - Directly used in Scribee project.
- Next step will focus on knowledge management
  - Need for an “usable” prototype to evaluate user experience and impact on users’ productivity,
  - We plan to study conversations in a larger, long term and interactive experiment (integration in email client software).

AT  
THE  
SPEED  
OF  
IDEAS™



[www.alcatel-lucent.com](http://www.alcatel-lucent.com)