# Abstract

Reaching out to a (?) large-scale of people via Internet is a fast and cost efficient way compared with postal mail or telephone. Therefore, email has been used not just for research, but also for marketing, customer support, and other data collection purposes. However, getting an acceptable response rate on the sent out emails requires additional efforts from the researchers’ side. This thesis investigates a communication system, which contributes to/towards increasing the response rate while minimizing the burden on the researchers’ side.

To achieve this, the system constructs a workflow supporting researchers to extract information, providing rule based automated decision making mechanism on respondents’ emails, and personalize the content of the emails with the respondents’ information which is extracted from the current state or earlier conversations. The System (or just: It) also provides an option to enable contribution of other researchers to interact with the workflow under the permission of the initial researcher. Therefore, distribution of the (this) work can ease individual’s efforts on the mass email communication. This feature can be further extended on enabling crowd workers on distribution of the work. What do you wanna say here? The crowd workers can participate in the work, they can get the results of the work or the work can be distributed to the crowd workers?

This thesis demonstrates that providing a proper workflow and the possibility of an assistant contribution, a mass email communication can be achieved as if each email is individually tailored to each recipient, which contributes to high response rates. Therefore, while it keeps the efforts on the creation of emails to a minimum, it maximizes the number of people communicated to.

# 1 Introduction

Increased Internet usage turned email into a tool for communication replacing telephone and regular mail (Norman and Lutz, 2000; Madden and Rainie, 2003). There are many use cases showing that email plays a huge role as a communication tool. Some of them include marketing for engaging clients, customer support for offering assistance after sale, surveying people to get their opinion on a topic, and many other cases showing that email become essential part of our daily life.

However, when the amount of people you want to reach increases, the way how you compose the emails and extract the information changes. BECAUSE The personal effort will not be enough anymore to individually tailor the emails according to each recipient or reading the respondent’s emails to extract the answers that you seek for. As a result, researchers tend to use online or software tools to send out generic emails to recipients with a non-adequate personalization, which is known as one of the important factor to increase response rates (Dillman, 1991; Schaefer and Dillman, 1998). Such emails are treated with low priority, which results low response rates at the end (Dillman, 2009, page 272).

There are several products in the market focusing on email communication and data collection. Customer relationship management (CRM) software keeps track of a company’s communication with their clients. Help desk software offers a platform to solve customers’ problems or provide guidance regarding products. Email marketing applications help out sending commercial messages to groups of people. Finally, survey applications aid to conduct online surveys to get people’s opinions and behavior. One of the common properties of all these applications is their dependency on email communication. However, none of these mentioned tools offers a complete workflow to help out a researcher to communicate by email with a great amount of people in a personalized manner and as easy as possible like communicating with an individual. (I changed it to avoid using the adjective possible two times in one sentence ;-))

The goal of this thesis is to understand the possible workflow of a personalized mass email communication, and to show that it is possible to reach a great amount of people by keeping the communication personalized at the same time. A complete system, named Myriad, has been developed to demonstrate the practical aspects of this idea.

## 1.1 Email as a data collection method

Nearly 600% growth rate on world-wide internet usage between 2000 to 2012 makes Europe’s 63% and North America’s 80% overall population internet usage proportion (Group, 2012). Email is ranked as the most popular online activity along with search engine usage with 92% of online adult users (Purcell, 2011). Also, the connectivity and the flexibility have been increased with the introduction of smart phones and tablet devices (Madden and Jones, 2008). In addition to these facts, email has low cost and quick turnover compared to regular mail or telephone communication (Zikmund and Babin, 2006). Therefore, email as a part of communication is considered as a viable option for data collection as well (Zikmund and Babin, 2006).

There are several reasons for data collection depending on the situation. However, purposes of data collection can be group under the following three categories (Sue and Ritter, 2011) (Babbie, 2012, pages 92–94):

1. To explore and get information about a topic

2. To describe the events and the situations

3. To explain things by questioning

To illustrate these purposes to see how we can use email to explore, describe, and explain things, let’s suppose that we have an online learning platform offering various courses publicly:

**Exploration** Offering online courses is a relatively new trend; therefore we do not have much previous knowledge about the topic. To explore the popularity of the platform, we need to ask the platform’s users questions about: Why are they attending our online courses? Have they taken any online courses before? What are their income levels? Figuring out the answers to these questions will help us to improve the system or to decide its future. For example, the aggregated answers to the income level question will make us decide whether to charge the users for their usage or offer it for free and find some sponsors to make it viable.

**Description** Our goal can be to describe characteristics of the online learning platform’s users. The questions helping us to describe this can be: Where do they come from? What are their age ranges? Have they attended a college? At the end, we might end up with a user profile like: users, at the age of 16 – 22, who have never attended to a college, and coming from less developed countries. Knowing our users’ portfolio according to this outcome can help us to attract organizations who have already had engagements to support those countries’ young population. Hence, they can leverage our platform as a tool to reach those populations.

**Explanation** We figured out that our platform’s users’ age range is between 16 -22 in our descriptive study. The reasons of why this ended up like that make our explanatory purpose. The questions like how often they are connected online or have they attended a college or a similar high level education institute might help us to find out the answer of why young people use our platform more frequently than older people. Collecting such statistics may help us to develop an explanation to a topic.

Since all of our registered users provided their email addresses as a primary and mandatory contact medium, we can use email to conduct our data collection whether the reason is to explore, describe or explain the user trends on our online learning platform.

## 1.2 Problem Statement

To date (what do this two words mean? ) , email as a popular medium for communication has many use cases including to reach groups of people (or: a group of people) to explore, describe, and explain things. However, when the group’s size gets larger, it becomes difficult from the researchers’ perspective to manage the state of the communication as in small groups. Therefore, researchers tend to write generic emails ignoring or using inadequate recipient specific information with the help of a software or online tool in the emails. This results low response rates since recipients become aware of being part of a large group, hence feeling less important and valued, as well as the chance to volunteer to reply the email gets less. On the other hand, if researchers individually tailor those emails according to recipients, it will require much more additional efforts and as a result costs, hence reducing the advantages of using email as a communication medium.

Even though, there are many solutions in the market to support email communication, there is no individual product allowing researchers to reach larger groups with minimum effort and keeping the communication personalized at the same time.

The main goal of this thesis is to show that personalized communication with large groups is possible when a proper workflow is provided. To achieve this goal:

1. Examine the workflow of an email communication with large groups and possible exceptional cases on this flow

2. Investigate the effects of email content’s personalization on the response rates

3. Describe how an adequate amount of personalized data can be supplied

4. Comparison of existing products claiming to provide solutions on email communication and collection of respondents’ information

5. Describe the design and implementation of an application satisfying the mentioned workflow to aid researchers including the initial prototype

6. Show how assistants can support the mentioned workflow

7. Real life use cases of the application and its users opinions about the application, and latest statistical information giving insight about how and in which way the application is used by its users.

This thesis also contributes on the following areas:

1. Email as a data collection method

2. Surveying with email

3. Defining a workflow on a mass email communication

4. Possible crowd sourced assistant usage

5. Personalization of email content

## 1.3 Outline

Outline goes here

# 2 Foundation and Related Work

This chapter presents the related work on the data collection domain. Even though, the technology is different for email surveys to collect data from well-established regular mail surveying methods, the nature of the communication is similar to self-administrated questionnaires (Schaefer and Dillman, 1998). Therefore, the chapter will also investigate the mail surveys in a way to emphases (I think this word I mistaken, couldn’t find the exact translation it’s used as a noun Emphasis in English but means sth else) the points which are also related with email communication, and the earlier studies on response rate influences.

## 2.1 Surveys and data collection

A Survey is defined as a system for collecting information (Sue and Ritter, 2011, page 3). It helps to learn about people’s opinions and behaviors (Dillman, 2009). The produced data during or at the completion of the survey belong to the data collection process. Therefore, data collection is a fundamental step to produce useful data to enable analyzes on researches (Groves et al., 2009, page 149). These researches include (but not limited to) many disciplines like sociology, statistics, psychology, marketing, economics, and heath sciences.

### 2.1.1 Email surveys

Comparing many different characteristics of surveys and interviews, the concerns regarding speed and cost make the most powerful differences (Sproull, 1986; Schaefer and Dillman, 1998). Email surveys offer more rapid surveying than other methods including regular mail and telephone surveys. In addition to that, email surveys are inexpensive since it removes the postage, paper and printing, and interview costs (Schaefer and Dillman, 1998).

Sproull (Sproull, 1986) identified the characteristics of email with an organizational research, within a (the? Instead if a, cause fortune 500 is a list) Fortune 500 office products and systems manufacturer, who were using email for 12 years in the organization and over 80 percent of all employees in the selected unit had (past cause you say: at the time of the research) email access at the time of the research. Selected candidates were separated into two groups. The data collection protocol within the organization asked each of the group’s participants series of questions regarding their 3-day old email inbox. Both groups filled out the questionnaire and answered open-ended questions either electronically or in writing.

The result of the study indicated that the average duration of data collection time for the email version was less than a week, which is half of the duration of the written version. While the response rate of the email version was 73 percent, the conventional written version’s rate was 87. The percentage of missing data in the questionnaires was .2 percent in the written version, and 1.4 in the email version. There were no differences in mean (? mean or main?) answers in email version comparing written questionnaire. (The main answers in the email version didn’t differ from those in the written questionnaire)

In another study from Sheehan and Hoy (Sheehan and Hoy, 2006), where they administered only an email survey to query individuals about their on-line behaviors and their attitudes and opinions regarding privacy, they have reached the shortest response time with 3.65 days comparing with earlier studies conducted until that time (See table 2.1). (you have to combine all these or change the first words, cause first you say in another study, then you say how they did the study, but to say what happened in the other study you started a new sentence)

In addition to speed of the email surveys, cost benefits have been indicated in Sheehan and Hoy’s study (Sheehan and Hoy, 2006) also concluded that email is an extremely cost-efficient method for data collection, where the total cost estimated at $470 ($30 for printing out the responses, $440 for 22 hours computer time to download surveys for printing) while postalmail is estimated at $6,500 (printing, postage, survey, and reminder mailing).

In another study from Mavis and Brocato (Mavis and Brocato, 1998), the email survey was nearly seven times cost efficient than postal survey. This includes labor hours, survey materials like booklets, mailing labels, envelopes, and postage costs. Total time spent into postal survey was 33 hours, but it only required 12 hours for the email survey. Final cost was $503.36 for postal survey, whose $305.36 was spent for postage part, and remaining $198 was spent for student labor cost. The only cost resulted from email survey was student labor cost, which was total $72.

Moreover, Paolo, et al., (Paolo, 2000) reported that people made longer open-ended response comments in email version of the survey compared to the mail version. While the average number of words per comment was 58.33% in the mail version it was 75.40% in the email version. Bachmann, et al., (Bachmann, 1999) had the same finding in 1995 and 1998, where open-ended questions were responded more likely by email recipients than mail recipients. In a latter study conducted in 1998, researches also found that email respondents were more likely to expand their answers, even it was not suggested by the survey, resulting in more candid responses than mail surveys. Responses to open-ended questions are one of the important measures to determine the quality of the returned surveys (Bachmann, 1999).

Given these advantages and positive benefits of email surveys, the next section will provide information about survey errors applicable to all type of surveys. (This is a comment for all the parts: without “the” it often sounds like slang to me, but it doesn’t need to be wrong ;-) just my personal feeling. So you can just verify it again and then forget my corrections on the “the” if they are not correct)

### 2.1.2 Survey Errors

Sample surveys are quantitative estimation of the distribution of a characteristic in a population by obtaining this information from a small portion of the corresponding population (Dillman, 1991). To generalize results from a small portion, which is a sample, to a population, following sources of errors needs to be considered (Dillman, 2006, page 9) (Dillman, 1991):

**Sampling Error** The more number of people surveyed, the larger degree of precision can be achieved. Therefore, the limitations on the number of people surveyed are considered under the sampling error. For example, while public opinion of 100 people results +**-**$10% of the true percent, 2,200 people results higher confidence with the percent of **+-**$2% (Dillman, 2006, page 9). The surveys relying on predefined list of recipients considered that the list is randomly generated or with a systematic sampling. Hence, it has got little research to reduce sampling errors comparing with face-to-face interviews in which multistage cluster designs1 are used due to cost and time limitations (Groves et al., 2009, page 106) (Dillman, 1991).

**1Cluster sampling selects preexisting groups of population elements instead of a single element of the population**

**(Groves et al., 2009, page 106). Departments of a university or households in a block represent**

**a cluster of people. When the allocation of those sampling resources are stratified and based on multiple**

**stages, frequently three stages, it is called multistage cluster sampling. First step selects the sample of**

**counties, followed by the blocks within those counties, and finally the dwellings from the chosen blocks**

**(Scott and Smith, 1969).**

**Coverage Error** When the list of surveyed people does not include all the elements of the population, coverage error happens (Dillman, 2006, page 9). Coverage error is considered one of the biggest issues of surveys since while surveying general public (Dillman, 1991).

**Measurement Error** When a respondent’s answer is hard to evaluate or cannot be compared with other respondent’s answers or there are inconsistencies between the observable variables like opinions, behaviors, or attributes and the survey responses, measurement error happens (Dillman, 2006, page 9) (Dillman, 1991). The possible reasons might depend on poor wording or order of the questions or the characteristics of the surveyed person such as incapability to provide correct answers or motivational factors (Dillman, 1991).

**Nonresponse Error** When there are large amount of people who do not response, and their characteristics are different from the ones who responded, then it results nonresponse error (Dillman, 2006, page 9). Low response has been considered a major problem, and many researches have focused on improving the response rates (Dillman, 1991).

## 2.2 Response Rate Influences

As mentioned in the previous section, one of the survey errors is the nonresponse error. Researchers have concerns regarding response rates, since responses coming from survey participants may be substantially different from those of nonrespondents, which will result in a biased estimate of representation of the population (Bogen, 1996).

Low response rate was even considered shortfall of the email methodology despite to its advantages (Bachmann, 1999). In table 2.1, there are nine studies where both postal mail and email are compared side by side. Out of those nine studies, four of them show high response rate on postal mail, three of them got higher response on email and two studies did not show any significant differences. Parker’s (1992) study of AT&T employees was the only study which got an acceptably high response rate by email. Schaefer and Dillman attributed this fact to the novelty of email and sent emails were carefully examined instead of considered company junk email (Schaefer and Dillman, 1998). Mavis and Brocato stated that studies cited by others in support of email surveys, also shown in table 2.1, did not compare email data collection with more traditional methods, and their study design and analyses varied greatly (Mavis and Brocato, 1998). Sheehan and Hoy also take the attention to many of these studies’ small and homogeneous population, therefore it may not represent larger population groups’ response tendencies (Sheehan and Hoy, 2006).

Therefore (two times therefore, maybe you can write: Because of that or sth.), researchers investigated on how to increase response rates at email communication. Schaefer and Dillman (1998) conclude that even though, the technology for email is quite different from well established postal mail surveying methods, the communication is considered similar to self-administrated questionnaires delivered by post. Hence, the techniques used to increase response rates on postal mail can be applied to develop an email methodology (Schaefer and Dillman, 1998). Following techniques are the ones where researchers focused on their effects on response rates.

### 2.2.1 Length

For many people the time required to spend on survey is considered the biggest cost (Dillman, 2009, page 26). The study from Heberlein and Baumgertner (1978) also states that the length of the survey has a negative effect on mail survey response rates, where they stated that each additional question reduces responses by .05% (Heberlein and Baumgartner, 1978). On the other hand, Bradburn (1978) suggests that the length of the survey is correlated with its importance, therefore it will increase the efforts both on researchers and respondents side resulting a higher response rate (Bradburn, 1978). Bogen (1996), in his literature review, concluded that the relationship between interview length and nonresponse is weak and inconsistent (Bogen, 1996).

### 2.2.2 Multiple Contacts

Researchers found that the number of attempts to contact people increases the response rates (Heberlein and Baumgartner, 1978; Schaefer and Dillman, 1998). The scenarios for multiple contacts include pre-notification contact, which is a brief notice for the main request, and follow-up contacts aiming to the people who did not respondent at the initial contact. Heberlein and Baumgertner (1978) showed that follow-up mailing has a mean return rate of 19.9% at the initial contact, and continued with 11.9% and 10.0% for the second and third contacts, respectively (Heberlein and Baumgartner, 1978). Schaefer and Dillman (1998) also stated the same conclusion for the multiple contacts for email in their literature research. According to this, the average response rate for email surveys with a single contact was 28.5% while 41% and 57% for two and more than two contacts, respectively (Schaefer and Dillman, 1998).

### 2.2.3 Personalization

Personalization has been addressed as an important factor to increase response rates by many researchers (Dillman, 1991; Schaefer and Dillman, 1998). It builds a connection between the respondent and researcher by making the respondent feel important, and drawing the respondent from out of the group (Dillman, 2009, page 272). Dillman and Frey (1974) conducted a study to see the effects of personalization, where they reached half of a university alumni sample via personalized cover letters, while the other half got impersonalized letters. The personalization treatment included personal salutations and real signatures on the mails. They achieved nearly 9% greater response rates for the personalized group (Dillman and Frey, 1974). It is also stated that this type of personalization techniques can be also applied to emails (Schaefer and Dillman, 1998). In the next section, we will continue with the applications of personalization in emails, and give the results of some studies.

## 2.3 Personalization of emails

Studies on mail surveys showed that personalization increases the response rates (Dillman, 1991; Schaefer and Dillman, 1998). Personalization is also important for email communication since it builds a connection between the respondent and researcher as in the mail surveys studies, and make them feel more important and valued (Dillman, 2009, page 272). With this argument, Dillman, et al., emphasized the social exchange theory2

**2Social exchange theory was considered as a frame of reference to other theories rather than a theory by**

**itself. It implies a two-sided, mutually contingent and rewarding transactions or exchanges (Emerson, 1976)**

On the other hand, Barron and Yechiam (2002) stressed on the socio psychological phenomenon, the diffusion of responsibility, which is also an outcome of volunteer’s dilemma. In the volunteer’s dilemma one player is needed to volunteer in order to reach the outcome preferred by all the others in the game. However, each person might be inclined to hoping that somebody else will volunteer, resulting in a higher utility of not volunteering than volunteering. According to this, the more people in the group size, the less probability of volunteering will result, which produce the diffusion of responsibility effect (Barron and Yechiam, 2002). In order to experiment the effect of diffusion of responsibility in the context of email requests, they sent emails asking for help either to single addresses or to a list of five addresses. In the email body (see Appendix A), a fictitious graduate student asked a question to know if the university has a biology faculty, whose answer was well known to anyone familiar with the institute. The result of the study showed that the proportion of replies where they used single email address in the "To" field got 20% higher response than the replies where they used groups of email addresses. In addition, the study qualified the given responses according to its helpful level, and the proportion of "very helpful" replies in the single email address condition was 187% higher than the groups of email addresses condition.

Another outcome regarding using multiple email addresses in "To" field resulted concerns from respondents in the study of Selm and Jankowski (2006). An introductory email including a link to a web-based questionnaire was sent to recipients to explore the opinions of elderly Internet users about an electronic political debate. One of the respondents remarked his concerns regarding the confidentiality when the header of the email contained all the email addresses of the other respondents explicitly. His reaction was quoted in the study as in listing 2.1.

Even though, the authors entitled that person as "skeptical" and his reaction as a "vivid skepticism", today it is one of the biggest concerns regarding email confidentiality, and it might result embarrassing situations from the research or the business perspective. A very recent email message (See listing 2.2 for the excerpt) dropped in my email inbox verify the importance of confidentiality.

In another study by Heerwegh (2005), personalization was applied to the salutations in the emails. The randomly drawn 2,540 samples from the student database of Katholieke Universiteit Leuven, Belgium were separated into equally sized two groups. In the non-personalized group, the salutation of "Dear student" was used, while in the personalized group "Dear [First name] [Last name]" was used. The email content was an invitation to a web survey which was about adolescent attitudes towards marriage and divorce. The result of the study showed that the personalization applied group got 6.9% higher login rate to the survey than the impersonalized group. Therefore, they concluded that increased response rates were in line with social exchange theory and with the diffusion of responsibility theory (Heerwegh, 2005).

In addition to personalization of salutations on the emails, Joinson and Reips (2007) stated the power of its combination with the power or status of the sender. In the study, a group of discussion panel students of Open University UK were sent an email invitation to complete a survey. Panel members were assigned to one of the conditions where salutation was modified in "Dear student", "Dear John Doe", and "Dear John". The sender power was manipulated on the first and last lines of the emails by assigning a neutral power saying that "From <name> (Strategy, Planning, and Partnerships), The Open University" and a high power "From Professor <name>, Pro-vice chancellor (Strategy, Planning, and Partnerships), The Open University". The results showed that the highest response rate was achieved when a personalized invitation came from a high power source and lowest when an impersonal one came from a neutral power source (See table 2.2). The possible reason for this was stated as personalized salutations increase people’s sense of identifiability, and its combination with a high power audience increase socially desirable, strategic behavior (Joinson and Reips, 2007).

As aforementioned studies show different forms of personalization increase the response rates in email communication. However, it has become very easy to add personalized information into email thanks to the software. Dillman, et al. (2009) stated that over-personalization using software tools might easily result impersonal messages, and gave an example (Dillman, 2009, page 237-238): "Dear Don Dillman, I am writing to inform you and your wife Joye that the XYZ Company has created a new dog food that we are sure your Boston Terrier, Crickett, will find to be very tasty. We would like to send a free sample to your home in Pullman, Washington." In this message, there is overwhelmed personalization with the usage of person’s wife, their dog’s type and name, and their home address. Moreover, experienced email users can identify if a message is written by a person or computer generated by looking appearance of one’s name in certain locations, and similar patterns for other information (Dillman, 2009, page 272). Therefore, it becomes difficult to have a correct amount and tone of personalization. The more daily interaction with digital devices will make the true authentic personalization rarer, hence achieving it will make it more important and effective (Dillman, 2009, page 238).

2.4 Conculusion

In conclusion, researchers conducted many studies regarding mail surveys as a data collection method, which has been more than any other survey methods (Dillman, 1991). Some of those studies tried to answer the question of nonresponse error, which has been considered one a major problem comparing with other survey errors as discussed in section 2.1.2. According to those mail survey studies, personalization has been addressed as an important factor to increase the response rates by many researchers in addition to other influences affecting response rates as identified in section 2.2. With the advance of world-wide internet usage, researchers has been started to consider email as a data collection method, because of its cost and speed benefits comparing with other data collection methods as discussed in section 2.1.1. However, some studies showed that response rates on email surveys are lower than regular mail surveys despite to its advantages; in addition, it may pose a burden to researchers during the collation of responses since email communication do not emphasize on any structure like in web forms or even respondents may come up with additional clarifying questions (Selm and Jankowski, 2006). Therefore, even the technology for email is different from mail surveying methods, researchers considered the response rate influences of mail surveys for email since the communication itself is the same. In section 2.3, several studies applying different type of personalization is mentioned. Some of those studies modified the header of the emails to study the diffusion of responsibility, other studies change the salutations and signatures of the emails, which resulted increase on response rates in emails. On the other hand, those studies did not consider the increased awareness of recipients to the possibility of computerized personalization techniques, which results over-personalized emails. As well as, none of the studies has taken the attention to the personal efforts of researcher while extracting information from respondents’ answers. This thesis will try to focus on those shortcomings on those studies as well, and provide a candidate solution to overcome those issues.