University of Economics, Prague

Faculty of Business Administration

International Executive MBA Programme (ISBM)

Master Thesis

The Transmission of Values and Culture in Open Source

Exploring the transmission of the "Apache Way" culture into Chinese initiated and non-Chinese initiated projects

by

[**Sharan Foga**] June 2019

Supervisors: Ing. Aleš Kubíček, Ph.D. & Justin McLean

Contents

List of Figures

Chapter 1

Introduction

1.1 Background

The Apache Software Foundation (ASF) is a 501(c)(3) charitable organisation that was founded in 1999 on a single open source project, called the Apache HTTP Server project. One of the key goals for creating a legal entity was to ensure that all ASF projects would continue to exist beyond the participation of the original project creators or individual volunteers. The type of entity selected, a 501(c)(3), was different from many other foundations because it was more aligned to the ASF mission of creating software for the public good.

The values, knowledge and governance model that developed as part of the creation of the Apache HTTP Server project is the source of the ASF culture also known as "The Apache Way".

The Apache Way is dependent on one central tenet – meritocracy, and this is embedded into all layers of the ASF. From formal governance, the election of directors and members, all the way through to the project communities and the recognition of individual contributions, merit plays a crucial part.

At the time of writing, the ASF was made up of over 350 projects and software initiatives, and it is claimed that each of these projects demonstrate and accept

"The Apache Way" as their cultural model.

In 2015, Apache Kylin became the first project initiated and contributed from China to successfully complete incubation at the ASF. It went on to graduate and become a top level project at the Apache Software Foundation.

As a US founded, yet global open source foundation, this was a significant event for the ASF. Why? Because a key part of the graduation process for becoming a top level project (TLP) involves understanding, acceptance and demonstration of ASF values and culture.

So what happens when a culturally Chinese project comes into contact with a different environment, culture, language and behavioural norms? In the case of Apache Kylin and other subsequent Chinese initiated projects, simply graduating as a TLP may not necessarily prove the long term acceptance or embedding of another culture.

This focus of this study was to explore four projects contributed from Chinese contributors that have undergone or are currently undergoing incubation at the ASF and to look for both qualitative and quantitative evidence of cultural embedding as a result of the Apache Software Foundation incubation process.

1.1.1 Defining Culture

We accept that it exists yet finding a single definition of culture has always been problematic. Perhaps it is simply created by the environment around us? (Hofstede et al, 2015).

Culture can influence our everyday decisions and behaviour because we carry our cultural identity with us as we interact with others. This means that the values we display will impact those around us. (Allen, 2003).

If culture is an unspoken communication (Guiso, Sapienza & Zingales, 2015) then it would rely on observing the cultural behaviour in action, but not being able to interact with it to understand the reasons behind the behaviour. In order to

fit in, one result could be to simply mimic what you see, and culture is more than copying the behaviour of others, there needs to be interaction.

If culture is something that develops in the social groups we engage with, then the fact that we are already engaged socially implies some commonly shared group characteristics (Seidler, 2010). So how can people who do not share these characteristics ever become involved or participate culturally?

If culture is about how we interact and share information with the people around us (Hofstede et al, 2015) then there must be a way of communicating and transmitting the key values and behaviours to others. We impact others with our cultural identity whenever we interact.(Allen, 2003)

The focus of this study was twelve ASF projects, four of which are Chinese initiated, that have been through incubation to see if any cultural changes occur as part of the project evolution and graduation to top level project.

1.2 Purpose

The purpose of the study was to document how the Apache Incubator process attempts to embed the open source culture called "The Apache Way" into new projects that appear to come from a different culture.

1.3 Research Questions

1. What evidence can we find of cultural embedding?;

2. How much of a cultural difference is there between the Chinese contributed projects and non-Chinese contributed projects?;

3. How successful is Apache Incubator in embedding ASF culture in the Chinese contributed projects?

1.4 Significance

This case study was used to investigate the process of embedding a culture using several projects at different stages of the incubation cycle. It uses a mixture of historic, retrospective and current real-time information in an attempt to capture the cultural evolution.

This work will help inform the Apache Software Foundation about its processes and the effectiveness of the Apache Incubator for cultural embedding.

Chapter 2

Literature Review

2.1 National Cultural Elements

Culture is the software of our minds that helps us to interact and share information with the people around us (Hofstede et al, 2015). The world is a very big place and not all of us were raised in the same environment, which means that we are culturally different. So using Hofstede's computer metaphor, our software versions may have compatibility issues to the extent that we don't understand or we misinterpret the cultural behaviour and signals of others.

The most well-known framework for defining cultural differences on a national level was developed by Geert Hofstede. His research highlighted six key areas or dimensions that could be used to assess the underlying structure of a culture. The six dimensions are:

1. Power Distance Index (PDI);
2. Individualism vs. Collectivism (IDV)
3. Masculinity vs. Femininity (MAS)
4. Uncertainty Avoidance Index (UAI)
5. Long Term Orientation vs Short Term Orientation (LTO)

6. Indulgence vs. Restraint (IND);

Each dimension describes a specific aspect where cultural differences could occur.

In this study, we are looking to initially identify cultural differences between China, as the source of the contributor initiated ASF projects, and the US where the Apache Software Foundation is legally registered.

If we look at Hofstede's data for China:

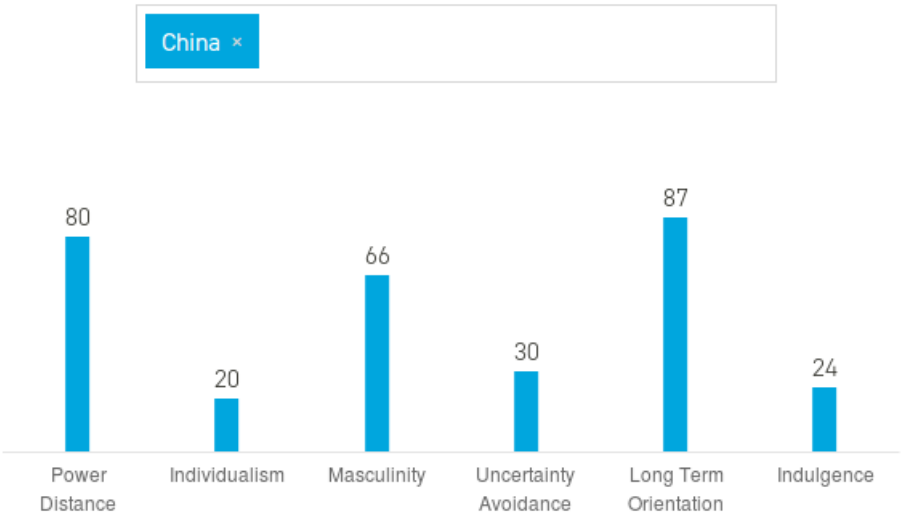

Figure 2.1: Hofstede Cultural Dimensions: China

And compare it to the US which is where the ASF was created as a foundation.

We see that the two most significant cultural differences are found in the dimensions of Power Distance (PDI) and Individualism (IDV). As well as the US, the members of the ASF founding group (the Apache Group) came from Germany, UK and Netherlands.

If we a look at the Hofstede cultural dimensions for these countries then we see a similar pattern to the US profile. So for this study, we will focus on these two dimensions PDI and IDV to describe the cultural differences within the Apache Software Foundation open source context.

Power Distance is focussed on indicating the equality of power or influence between individuals within a society. It also highlights their acceptance of the dif-

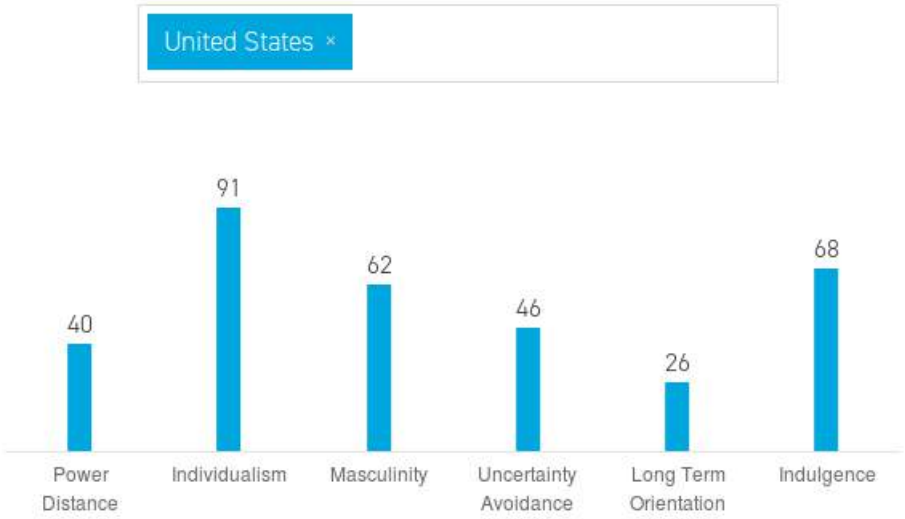

Figure 2.2: Hofstede Cultural Dimensions: United States

Figure 2.3: Hofstede Cultural Dimensions: Germany, Netherlands, United Kingdom

ference. In short, how comfortable are you at accepting that other people within the society have the power to influence your life. A high score highlights that people generally expect and accept power inequality, whereas a low score indicates the opposite.

As a country, China scores high in Power Distance (80) indicating that it is a society that believes that inequalities amongst people are expected and acceptable. Compare this with the US score of 40. So here we can see that the Chinese cultural

expectation is to have little or no power to influence, yet the US cultural expectation is the opposite, to have the power to influence.

If we go back to the previously mentioned ASF cultural tenet of meritocracy that gives everyone the power to potentially influence, then we can see that this value is more aligned with the US score.

Societies can be individual, where people focus on being independent, looking after themselves with little or no reliance on others, or collective, where there is a strong group focus and the needs of the group outweigh any individual need.

As a country, China scores low in Individualism (20) indicating that it is a more collectivist society based on a group culture rather than individualist. Compare this with the US score of 91. So here we can see that the Chinese cultural expectation is to be group based with common goals rather than focussed on the any one individual, yet the US cultural expectation is that individuals are more important than the group.

Going back to the way the ASF organises its projects, this is done around communities and it is the community that ensures the longevity of a project. Based on this we can see that this is more aligned with collectivist thinking and therefore the Chinese score.

This seems to highlight a paradox in the sense that "The Apache Way" is not fully aligned with US culture even though it was founded in the US but consists of aspects of both US and Chinese culture.

2.2 Open Source Culture

Does open source culture really exist? The phrase "open source" has been used to describe politics, journalism, biology, music and education (Ceraso & Pruchnic, 2011). So why would developing software and making it available for free suddenly make it cultural? Open source culture could simply be another popular buzzword with little or no substance behind it.

Instead of a distinct culture, could it be that open source software development is simply a way to make use of the available work capacity of the people who volunteer to provide their time and energy for free (Ceraso & Pruchnic, 2011). If open source is purely focussed on writing and producing code then that alone isn't enough to define the development process as a culture.

Could it be that as people are used to being organised and managed in a work environment, in their participation in open source they have converted the terminology 'work culture' into 'open source culture' to describe their unpaid working environment? If it is only about producing software that also happens to be open source, then that isn't anything new as software development has been around since the 1960s and no one ever called it culture so why is open source any different?

In 1984, Richard Stallman founded the Free Software Foundation (FSF). Its very first goal was to create the GNU license. Some regarded this action as politically motivated (Kelty, 2004) and this focus on intellectual property rights promoted the ideology that something needed to be fought for.(Carillo & Okoli, 2009)

The GNU licence provided a legal definition as well as protection for free and open source software. Also the FSF slogan "Free as in Speech, not as in Beer" provided a clear link between free and open source software and the right to be heard.

Stallman himself, made the clear distinction between open source and free software highlighting that open source was a development methodology while free software was a social movement.(Zhou, Y. 2011) Here we have a clear distinction being made where only one is described as being social, and social interaction is one of the key aspects of communicating culture.

In 1991, Linus Torvalds created the Linux kernel and at the time it was seen to have been motivated as a challenge to the existing commercial market of proprietary software (Kelby, 2004). Having an open source software alternative for an operating system was significant and revolutionary. This event affected and impacted society. Once again we have a social movement but this time it is competitive and targeted toward the existing software corporations.

During the mid-1990s the most popular web server software in the world (the HTTP daemon) was free and open source. It had been developed by Rob McCool at the National Center for Supercomputing Applications (NCSA), University of Illinois. When Rob left the NCSA, the development and maintenance of the HTTP daemon stopped.

Many people were still actively using it so rather than abandoning the project, a small group of users decided to work together to provide fixes and extensions. This group became known as the "Apache Group" and as a result of their collaboration, less than a year later, they released their own version of the HTTP daemon and called it the Apache HTTP Server or the "Apache web server" as it is more popularly known. Again we see social interactions being used as the basis for creating a way of working and collaborating.

Although each of these open source movements were triggered by different events, there is a common factor of social links being an important factor. The triggers and environment will therefore have an effect on the resulting behaviour, values and communication mechanisms. And it is these that are commonly used to describe and define culture.

2.3 ASF Culture, Vision and Values

As mentioned previously, many open source projects are created as foundations or non-profit entities. Although some foundations are focussed on delivering benefit to their shareholders or members, the mission of non-profits is focussed around delivering benefits to the public. To recruit new contributors, to help achieve their mission open source projects need to provide something unique to attract contributors and encourage them to participate. Culture can be used as a way of attracting and retaining contributors.

The ASF has its own unique culture called "the Apache Way". It is a culture that was created directly as a result of establishing the Apache HTTP Server project and it has evolved over time.

When the ASF was established in 1999, a key goal was to ensure that all ASF projects would continue to exist beyond the participation of individual volunteers. This situation was directly related to the NCSA HTTP daemon where the original developer left the project and the Apache Group worked together to revive, maintain and successfully evolve it.

The ASF is based on the meritocratic model, where the contributions an individual makes are recognised. Each contribution earns "merit", and as an individual accumulates merit, they are rewarded by being allowed to take on more responsibility. Another key concept is that merit never expires, this means that a contributor will always keep the recognition of the work they have done.

Meritocracy has also been referred to as a "do-ocracy" because those that do more can achieve more responsibility. This means that the power to influence a project lies with the people that are actively contributing.

The ASF guidelines state that:

"individuals comprise the ASF"

For those who are new to ASF culture this concept could confuse because the reference to individuals is not promoting individualism. Instead it means that each contributor not only forms part of the ASF but also that they are recognised for it.

A key point to note is that merit is only recognised for individuals not companies which means that companies cannot gain merit and the power to influence. This helps keep the ASF a vendor neutral environment.

The ASF also promotes:

"Community over code"

meaning that the welfare of the project community is more important to maintain than producing software. For projects whose goal it is to produce usable software for the public good, it seems unusual to promote that software development is not the

main priority. It is the community that maintains the project so having a thriving and active community helps ensure a project will survive into the future.

2.3.1 The Apache Way in Practice

Assumptions can be described as unconscious behaviours and values that are deeply embedded in an organisation. These behaviours occur so naturally to the people who are part of the organisational culture that they may not even be aware of them. In many cases, only people outside that are not already connected with the culture can identify or recognise when these assumptions are being displayed.

All open source projects are different, and it will be these inherent assumptions that will form the core of each project's culture. The following are some of the observed potentially inherent assumptions of the Apache Way:

- Consensus
  - Decisions are made by consensus.
  - Consensus is not voting, it is showing an opinion, support or is used as a way of creating a path to move forward after disagreement.
  - Anyone that is part of an ASF community has the power to influence.
  - Getting consensus from a community is better than voting and reduces the adoption of "them" vs. "us" behaviour.
- Community Self Correction
  - Each project community is independent in the way it is allowed to develop and maintain itself.
  - Allow communities to self-correct if something negative occurs.
  - Intervention from the Board of Directors is a last resort solution.
- Community over Code
  - It is more important to have a collaborative community than it is to have the best software and a dysfunctional community.

  - Respectful interactions and collaboration can still occur even if both sides have opposite views.

- Different Hats
  - The understanding of different perspectives.
  - Commercial vs. Employee perspective.
  - Community vs. Individual perspective
  - Different roles within the ASF and need to clarify which "hat" is being worn in a particular situation (e.g Board Member, PMC chair, Member, Committer, Individual).

- Mailing List as the Source of Truth
  - If it is not on the mailing list, it didn't happen!.
  - Radical transparency - discussions are open and include the whole community.
  - Decisions need to be made openly.

- Merit
  - All individuals are capable of gaining merit.
  - Recognition of merit
  - Merit never expires

2.4 Chinese Culture and Open Source

East Asian cultures tend to say little and not display emotions openly, which differs significantly from the Western or European based cultures that have no fear in expressing what they think.(Caldwell-Harris et al. 2013)

Chinese culture values preserving group harmony. Any displays of negative emotions such as anger or criticism would act as disruption to that harmony. (Caldwell-Harris et al. 2013) Disharmony affects the ability of the group to work together

efficiently. In a collectivist society where the needs of the group outweigh the individual, displaying disrupting emotions might be avoided because it could centralise the focus on the individual rather than the group.

Being direct and perhaps even overly expressive is not unusual in American culture. But in Chinese culture, this tends to be the opposite and the real skill is actually about being indirect. (Caldwell-Harris et al. 2013)

The triggers and events that led to the initiation of open source through Stallman, Torvalds or the ASF didn't happen in the same way or at the same level in China. It is thought that the interest in open source and Linux was brought to China in the early 1990s by technical students and open source enthusiasts. (Murray, 2006).

As well as individuals, the Chinese government took on a sponsorship role to help promote open source (Pan & Bonk, 2007). One of their main aims was to help reduce the reliance on proprietary software and to use open source as a catalyst to kick start development in their own internal software industry (Zhou, 2011).

Limited availability of open source software and a lack of mature open source communities gave academic institutions a key role. (Pan & Bonk, 2007).They encouraged their students to focus on contributing to open source projects or to create new open source projects of their own.

1999 was a significant year which saw the founding of the first locally based Chinese Linux company – Xteam. (Murray, 2006). This caught the attention of other Linux software providers such as Turbolinux who entered the Chinese open source market with a Chinese language version of their product.

With government backing, in 2001, the Software Research Institute of the Chinese Academy of Science successfully released and distributed their own open source operating system called Red Flag Linux.(Pan & Bonk, 2007). Red Flag Linux was eventually discontinued in 2014.

With the initial and continued introduction of several open source products during the late 1990s and early 2000s, open source was brought to the forefront in China

both in the business and academic world.

Chinese open source users, initially started out as consumers or end users of open source software (Zhou. 2011) using the tools necessary to meet their needs. Now the situation appears to be changing and people are looking to contribute something back to the open source projects that have helped them.

The trend of giving something back seems to be growing amongst the Chinese open source community. As at 30th April 2019, the ASF had 8 top level projects;

- Apache CarbonData
- Apache Eagle
- Apache Griffin
- Apache HAWQ
- Apache Kylin
- Apache RocketMQ
- Apache ServiceComb
- Apache Skywalking

and 6 incubating projects

- Apache brpc
- Apache Doris
- Apache Dubbo
- Apache ECharts
- Apache Sharding-Sphere
- Apache Weex

all of which have been initiated from China.

2.5 Cultural Change

Hofstede's research identified certain aspects of national culture. This approach meant that culture was seen as something that could be packaged and referenced at the country level (Rose, 2007).

Technology has changed the way we interact, meaning that our social cultural interactions have changed too. We live in a virtual and electronic world that is no longer constrained by national boundaries (Rose, 2007). We can access and interact with culture from anywhere.

If culture cannot be transmitted (Rose, 2007) then how can it survive long term? And more importantly, how can it move from generation to generation and still exhibit the same social behaviours and traits?

The answer may be in the terminology we use. Maybe it's not about the transmission of culture but the relationships that it creates. We already know that culture is interactive and social, so perhaps culture is simply a series of relationships with common ways of seeing and doing things (Rose, 2007).

If culture is something that we need to internalise, then we will be exhibiting it unconsciously in everything we do and in every interaction that we have with others. (Rose, 2007) This means that we effectively have no control over when and how we transmit our culture to others. So by simply interacting we are transmitting something.

The group of individuals that started the "Apache Group" are still involved in ASF today. This means that there is still a direct and existing link to the original source of Apache culture. Being able to ask for clarification around the understanding of cultural values and associated behaviours can help to ensure that the culture is not deviating too much from its original form over time. In its simplest form, this is mentoring.

It is fairly straightforward to mentor a person, but what happens when you need to mentor a whole project and its community, or even 50 different projects? The

ASF has already asked this question and its response is Apache Incubator.

2.6 Apache Incubator

As the number of open source projects under its umbrella began to grow, the ASF as an organisation was created to provide governance and a legal framework for its projects. The Apache Incubator was created in 2002 with the role to guide and mentor projects wishing to join the Apache Software Foundation specifically:

"accepting new products into the Foundation, providing guidance and support to help each new product engender their own collaborative community, educating new developers in the philosophy and guidelines for collaborative development as defined by the members of the Foundation, and proposing to the board the promotion of such products to independent PMC status once their community has reached maturity"

It holds the responsibility for educating potential projects in the way the ASF would like their projects to behave.

To support new projects entering incubation, the ASF has developed a formal governance process, roles and responsibilities and a project maturity model. The project maturity model is optional although it is a useful framework and tool that can help projects benchmark themselves to see which aspects they need to focus on.

2.6.1 Roles and Responsibilities

At the highest level of governance is the ASF Board of Directors. It is they who approve the establishment of a new top level project, appoint the project chair and approve the project management committee (PMC) members. In order to reduce the risk of rejection, the Apache incubation process assesses each project to ensure it meets the minimum standard needed to be considered for graduation.

Reporting to the ASF Board is the Incubator Project Management Committee (IPMC). It is made up of people interested in helping or mentoring potential new

ASF projects. The IPMC's role is to accept new projects for incubation and then work with the project and its mentors towards graduation to top level project.

The process to accept a new project starts with creating an incubation proposal. Until the project is accepted, it is referred to as a "Project Proposal". Each Project Proposal needs a "Champion". This is generally someone who is already deeply involved with the ASF that can see value in bringing the project into incubation.

The Champion works with the proposed project to create the proposal. The proposal details the current status of the project and assesses the project aspirations for the future. It looks at why the project wants to become part of the ASF and discusses whether or not it would be a good fit.

Each Project Proposal also needs Mentors, to help them understand ASF policies and guidelines as well as ensuring that the communities develop the practices and culture advocated by the ASF. Mentors must be members of the Incubator Project Management Committee (IPMC) and generally 3 or more mentors are required.

Mentors provide advice, are active on the project mailing lists and can act as facilitators in case of disputes. The role of mentors is crucial in helping embed ASF culture.

2.6.2 Apache Project Maturity Model

As part of the collaboration process, an Apache Project Maturity Model has been developed. It is an optional framework model that can be used to assist in the process of evaluation for graduation and also provides transparency to projects on the areas being assessed.

During incubation the project will report to the ASF Board on a regular basis. Before submitting the report, it is reviewed by the project mentors who can comment on any areas before giving their approval.

The report can focus on the status of any of the following areas:

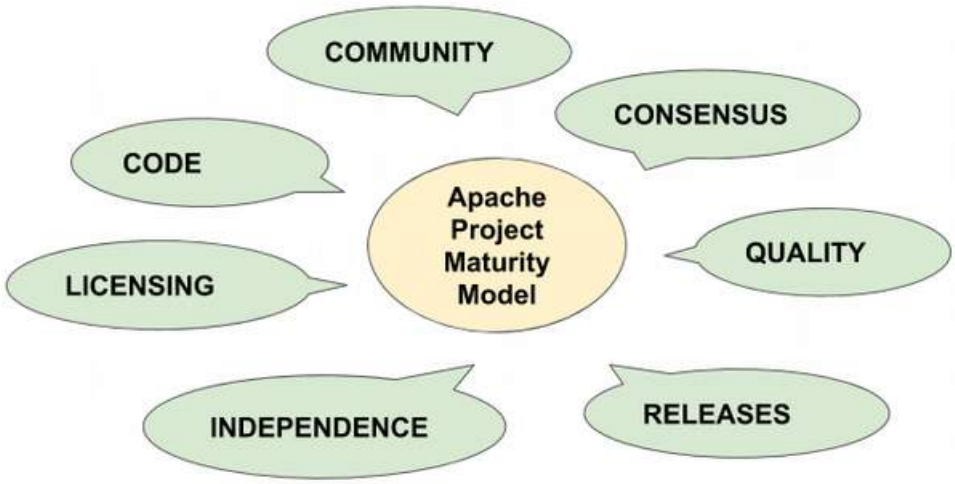

Figure 2.4: Apache Project Maturity Model

Each area can be further broken down into concrete tasks that incubating projects need to ensure are completed during their incubation.

Apache Project Maturity Model Detail Level

CODE	
Ref	Description
CD10	The project produces Open Source software, for distribution to the public at no charge.
CD20	The project's code is easily discoverable and publicly accessible.
CD30	The code can be built in a reproducible way using widely available standard tools.
CD40	The full history of the project's code is available via a source code control system, in a way that allows any released version to be recreated.
CD50	The provenance of each line of code is established via the source code control system, in a reliable way based on strong authentication of the committer. When third-party contributions are committed, commit messages provide reliable information about the code provenance.

LICENCING AND COPYRIGHT	
Ref	Description
LC10	The code is released under the Apache License, version 2.0.
LC20	Libraries that are mandatory dependencies of the project's code do not create more restrictions than the Apache License does.
LC30	The libraries mentioned in LC20 are available as Open Source software.
LC40	Committers are bound by an Individual Contributor Agreement (the "Apache iCLA") that defines which code they are allowed to commit and how they need to identify code that is not their own.
LC50	The copyright ownership of everything that the project produces is clearly defined and documented.

RELEASES	
Ref	Description
RE10	Releases consist of source code, distributed using standard and open archive formats that are expected to stay readable in the long term.
RE20	Releases are approved by the project's PMC (see CS10), in order to make them an act of the Foundation.
RE30	Releases are signed and/or distributed along with digests that can be reliably used to validate the downloaded archives.
RE40	Convenience binaries can be distributed alongside source code but they are not Apache Releases – they are just a convenience provided with no guarantee.
RE50	The release process is documented and repeatable to the extent that someone new to the project is able to independently generate the complete set of artifacts required for a release.

QUALITY	
Ref	Description
QU10	The project is open and honest about the quality of its code. Various levels of quality and maturity for various modules are natural and acceptable as long as they are clearly communicated.
QU20	The project puts a very high priority on producing secure software.
QU30	The project provides a well-documented, secure and private channel to report security issues, along with a documented way of responding to them.
QU40	The project puts a high priority on backwards compatibility and aims to document any incompatible changes and provide tools and documentation to help users transition to new features.
QU50	The project strives to respond to documented bug reports in a timely manner.

COMMUNITY	
Ref	Description
CO10	The project has a well-known homepage that points to all the information required to operate according to this maturity model.
CO20	The community welcomes contributions from anyone who acts in good faith and in a respectful manner and adds value to the project.
CO30	Contributions include not only source code, but also documentation, constructive bug reports, constructive discussions, marketing and generally anything that adds value to the project.
CO40	The community strives to be meritocratic and over time aims to give more rights and responsibilities to contributors who add value to the project.
CO50	The way in which contributors can be granted more rights such as commit access or decision power is clearly documented and is the same for all contributors.
CO60	The community operates based on consensus of its members (see CS10) who have decision power. Dictators, benevolent or not, are not welcome in Apache projects.
CO70	The project strives to answer user questions in a timely manner.

INDEPENDENCE	
Ref	Description
IN10	The project is independent from any corporate or organizational influence.
IN20	Contributors act as themselves as opposed to representatives of a corporation or organization.

CONSENSUS	
Ref	Description
CS10	The project maintains a public list of its contributors who have decision power – the project's PMC (Project Management Committee) consists of those contributors.
CS20	Decisions are made by consensus among PMC members and are documented on the project's main communications channel. Community opinions are taken into account but the PMC has the final word if needed.
CS30	Documented voting rules are used to build consensus when discussion is not sufficient.
CS40	In Apache projects, vetoes are only valid for code commits and are justified by a technical explanation, as per the Apache voting rules defined in CS30.
CS50	All "important" discussions happen asynchronously in written form on the project's main communications channel. Offline, face-to-face or private discussions 11 that affect the project are also documented on that channel.

New projects entering incubation will need to report monthly to the ASF board for the first three months. After this it changes to once every three months. The report generally includes information about what the projects have accomplished and what things they are planning to work on towards achieving graduation.

This is where the Apache Project Maturity model can be used as a checklist to highlight any potential gaps. Although optional the model provides clear guidance for the targets projects need to achieve.

Each project community is unique and so may have different challenges to overcome which can take time, so there are generally no time constraints on incubating projects as long as they are active and progress is being made. In the case where project activity is very low the ASF board will request that at least 3 members of the project management committee confirm that they are still active in the project to remain to incubation.

Projects can also withdraw from incubation at the ASF. This is can be done for a variety of reasons, the main one being a lack of cultural fit.

Before any project can be considered for graduation it must prove that it

- has a community built around it that is active
- is a good fit for the ASF
- understands and uses the Apache Way culture
- has an existing project codebase
- understands the incubating process and the expectations of it

As well as legal and infrastructure requirements, there is a strong need for the project requesting graduation to demonstrate the following:

- An active and diverse development community
- The adoption of merit based community interactions
- Not being dependent on any single contributor and that there are at least 3 independent committers
- No single company or entity is vital to the success of the project
- ASF style consensus is been adopted and is standard practice
- The ability to tolerate and resolve conflict within the community
- Be able to create and execute a release in public that has been driven by community effort

If a project can fulfill these requirements and the Champion and Mentors support it, then a project can create a resolution to the ASF board to graduate and be formally established as an ASF top level project.

Chapter 3

Methodology

3.1 Methodology and Hypotheses

This paper will analyse the main elements of the ASF culture ("the Apache Way") and attempt to provide evidence to show the extent that this culture can be successfully transmitted to new projects. It will test and explore the following hypotheses:

1. As the original source of ASF culture, data from the Apache HTTP Server project can be extracted and used to create a cultural model
2. This cultural model can be used as a baseline to compare against other ASF projects see if they exhibit a similar cultural profile
3. This cultural model can be used as a baseline to compare ASF projects that have undergone or bypassed incubation in Apache Incubator to see if they exhibit a similar cultural profile
4. The projects contributed from Chinese contributors may undergo a more significant culture change than other non-Chinese related projects

Based on the previous assumptions, this paper will focus on the following:

1. Define the main concepts and values of the ASF cultural model using publicly available information

2. Define the role and processes of the Apache Incubator

3. Mine the public data available from the ASF code repositories and mailing lists for the Apache HTTP Server project to create a cultural baseline

4. Mine the public data available from the ASF code repositories and mailing lists for a series of Apache projects that were created after the Apache HTTP Server project to look for data indicators that validate or disprove the demonstration of elements found in the Apache HTTP Server cultural baseline

5. Mine the public data available from the ASF code repositories and mailing lists for a series of Apache projects that:

    - Have bypassed the Apache Incubation Process (i.e. no formal embedding of ASF culture)
    - Are in incubation or have graduated from Apache Incubator that were contributed from non-Chinese contributors (i.e. formal embedding of ASF culture)
    - Are in incubation or have graduated from Apache Incubator that were contributed from Chinese contributors (i.e formal embedding of ASF culture)

The resulting analysis will be assessed and compared against each other, and used to respond to the research questions.

3.2 Research Design

The purpose of this study was to gain a better understanding of the open source culture at Apache Software Foundation and to analyse a project that has a different cultural profile to see if cultural changes occur as part of the project evolution and graduation to Top Level Project. A key part of graduating and becoming a Top Level project involves the demonstration of Apache behaviour and cultural values.

The study focuses on twelve Apache projects, four of which are Chinese initiated, that have been through incubation at the Apache Software Foundation, to see if any cultural changes occur during the incubation process.

In order to have some comparative data, the twelve projects also include:

- four Apache projects which have undergone incubation that were initiated by non-Chinese contributors
- four Apache projects which did not undergo incubation that were initiated by non-Chinese contributors

The time period for the study was 1st January 2014 to 30th April 2019.

3.3 Environment

All ASF projects have publicly archived mailing lists. They are a legacy communication medium from the creation of the ASF that now forms an integral part of any ASF project.

While this style of communication may seem unconventional given the changing technology in social and communication media, an essential part of ASF governance is to ensure there is a record of community discussion and decisions.

Mailing lists allow for asynchronous communication so that people in different time-zones, or those that work on projects part-time or out of general office hours, are given both time and opportunity to join a conversation.

3.4 Research Tools and Indicators

The research will be carried out using the following tools, formulas and indicators.

3.4.1 Apache Kibble

Apache Kibble is a suite of tools for collecting, aggregating and visualising data and activity in software projects.

The following Kibble indicators will be used:

3.4.2 Pony Factor

Pony Factor (PF) measures the diversity of a project based on the contributions from individual contributors. It can be defined as:

"The lowest number of contributors whose total contribution makes up the majority"

of whatever is being measured (e.g. lines of code written, number of messages sent etc.).

Mathematically it is written as follows:

$$\sum_{i=1}^{P} C_n >= K.V$$

Where:

- P is the Pony Factor
- Cn is the number of contributions made by contributor n sorted by descending number of contributions
- K is the percentage of the total contributions we are looking for
- V is the total volume of contributions made

A higher Pony Factor means that a project has a good tolerance for continuing to survive if one or more of the core contributors leaves.

NOTE: Pony Factor includes all contributions from contributors whether they are still active or not.

3.4.3 Augmented Pony Factor

The Augmented Pony Factor (APF) is an adjustment to the standard Pony Factor calculation where contributions from contributors that are no longer active are omitted.

$$\sum_{i=1}^{P} C_n >= K.V - Contributions\ from\ non-active\ contributors$$

NOTE: The Augmented Pony Factor will not be used as part of the assessment but a description of it has been included here for completeness.

3.4.4 Meta Pony Factor

The Meta Pony Factor calculation is a work in progress. It attempts to measure the affiliation of a contributor based on the email address linked to the contribution. If developed further then this could help identify distinct organisations that are contributing.

3.4.5 Example of Pony, Augmented and Meta Pony Factors

The following example can help explain the use of Pony Factor, Augmented Pony Factor and Meta Pony Factor.

An open source project has 35 active contributors:

- Contributor A has made 19% of the contributions and works for Company X.
- Contributor B has made 15% of the contributions and works for Company X.

- Contributor C has made 22% of the contributions is no longer active and works for Company Y.
- Contributor D has made 12% of the contributions,and works for Company Z.
- Contributor E has made 9% of the contributions has no company affiliation.
- All other contributions were made the other 30 other contributors.

The Pony Factor would be 3 because contributions from Contributors A,B and C make up 56% of the project.

The Augmented Pony Factor would be 4 because Contributor C is no longer active so the 22% that they contributed would not be included as part of the calculation. This means that contributions from Contributors A, B,D and E make up 55% (the majority) of the project.

The Meta Pony Factor is 4 because we have 4 visible affiliations (Company X,Y, Z and no affiliation) associated with the Pony Factor contributions.

Relevance for the ASF

ASF projects are made up of individual contributors. Some of these individuals contribute to the software through code and some via non-coding or knowledge (e.g. documentation, testing, marketing etc.). The Pony Factors can measure the diversity of a project in terms of the division of work among the contributors. The higher the Pony Factor the stronger the tolerance for continuing to survive if one or more of the core contributors leaves the project.

Pony Factors are available for the codebase, email and the issue trackers. The Pony Factor of the codebase will focus on coding contributions whereas the email Pony Factor will incorporate non-coding contributions such as community participation and involvement via the mailing lists.

The Apache HTTP Server project was created directly as a result of a core contributor leaving the NCSA project so a key part of ASF culture is ensuring that projects can survive even when main contributors leave.

The Pony Factors can also help indicate:

- the growth of a project
- new contributors are being accepted
- merit is being rewarded by committer status

3.4.6 Sentiment Analysis

Sentiment analysis is a form of text mining used in the prediction of the emotional state of a person based on what they have written (Pang & Lee, 2008). This means that it can be used as an indicator to gauge people's opinions and reactions to certain ideas.

Data is collected in the form of text and an algorithm is used to identify keywords associated with an emotion. Any communication can be linked to several emotions so weightings are used to highlight the strength of the sentiment.

Comparative and relative sentiment analysis is a measure of how representative a mood being expressed is uniformly at the same or a similar level of intensity throughout a group or organisation. Apache Kibble includes a feature that compares the sentiment across the entire organisation to highlight how representative it is.

All ASF projects have public mailing lists as the main method of communication. Everything that occurs in a community will be reflected in the text interactions being recorded on the mailing lists. This means that running sentient analysis over the mailing lists can give an indication of the emotional state of a community.

The following mood sets will be used for the analysis:

- moods_good = set(['trust', 'joy', 'confident', 'positive'])
- moods_bad = set(['sadness', 'anger', 'disgust', 'fear', 'negative'])
- mood_neutral = set(['anticipation', 'surprise', 'tentative', 'analytical','neutral'])

3.4.7 Key Phrase Extraction

Key Phrase Extraction (KPE) is a method where key phrases or words are extracted that can summarise the main ideas or themes of a text or document. It has been successfully used for indexing journals and online content but for the purposes of this paper it will be used to extract any text that could indicate cultural ideas or language.

In this study we will use Apache Kibble's key phrase extraction function to extract key phrases from the mailing lists of the 12 projects we have selected for the study, to see if there is any indication of cultural ideas or language.

If a culture is embedded in a community then the language used should be indicative of that culture. The area known as 'cultural linguistics' attempts to research the relationship between culture and the language used (Sharifian, 2017). The expectation is that words and phrases that have a strong significance with ASF culture will be present.

3.4.8 Contributor Retention

Contributor retention is related to how successful a project is at attracting and retaining contributors. It can be broken down into the following areas:

- Active Contributors
  - How many contributors are active within the project?
  - Are contributors contributing regularly and remaining active over a longer timespan?
- Retained Contributors
  - How long contributors have been contributing?
  - The longer contributors have been retained, the more successful a project is at retaining them

- Contributors that have Left
  - How many contributors are leaving
  - The rate at which contributors are leaving
- Past Contributors that have Returned
  - How many contributors have contributed in the past and have returned to rejoin the community
  - The rate at which past contributors are returning

Contributor retention is a very important metric because as well as activity or lack of activity, it also shows if a project has a good mix of active contributors, is accepting new contributors or is mainly dependent on experienced ones.

3.5 Data Collection

All ASF projects have publicly archived mailing lists. They are a legacy communication medium from the creation of the ASF that now forms an integral part of any ASF project. It is the heart of a project and a place where people interact, communicate, collaborate, argue, agree and disagree. This means that it is an appropriate place to mine data to look for cultural indicators.

Data to create the cultural baseline will be extracted from the Apache HTTP Server project. The following data will be used:

- Apache HTTP Server Mailing List Archives 1996 – April 2019
- Apache HTTP Server Code Repositories 1996 – April 2019
- Apache list of selected projects Mailing List Archives 2014 – April 2019
- Apache list of selected projects Code Repositories 2014 – April 2019

The following 12 projects have been selected for comparison because:

- They have all joined the ASF in the last 5 years (from 2014-2019)
- They all have data available in Apache Kibble
- None of them existed at the time that Apache HTTP Server was created
- The culture they exhibit would have been created after the ASF was established

List of Projects Included in this Case Study

Project Name	Status as at 30th April 2019	Reason for Data Inclusion
Apache Kylin	TLP	Chinese Initiated Incubated Project
Apache Skywalking	TLP	Chinese Initiated Incubated Project
Apache Dubbo	Incubating	Chinese Initiated Incubated Project
Apache Weex	Incubating	Chinese Initiated Incubated Project
Apache Netbeans	TLP	Non-Chinese Initiated Incubated Project
Apache Beam	TLP	Non-Chinese Initiated Incubated Project
Apache Kudu	TLP	Non-Chinese Initiated Incubated Project
Apache Kibble	TLP	Non-Chinese Initiated non-Incubated Project
Apache Whimsical	TLP	Non-Chinese Initiated non-Incubated Project
Apache Bahir	TLP	Non-Chinese Initiated non-Incubated Project
Apache ORC	TLP	Non-Chinese Initiated non-Incubated Project

The projects can be broken down into three sub groups as follows:

3.5.1 Chinese Initiated Incubated Projects

Project Name	Started Incubation	Graduated Incubation	Length of Time in Incubation	Years as TLP
Apache Kylin	25-11-2014	18-11-2015	12 months	3
Apache Skywalking	08-12-2017	17-04-2019	16 months	0
Apache Dubbo	16-02-2018	-	14 months	-
Apache Weex	30-11-2016	-	29 months	-

3.5.2 Non-Chinese Initiated Incubated Projects

Project Name	Started Incubation	Graduated Incubation	Length of Time in Incubation	Years as TLP
Apache Netbeans	01-10-2016	17-04-2019	30 months	0
Apache Fineract	15-12-2015	19-04-2017	16 months	2
Apache Beam	01-10-2016	21-12-2016	10 months	2
Apache Kudu	03-12-2015	20-07-2016	7 months	2

3.5.3 Non-Chinese non-Incubated Projects

Project Name	Date Created as TLP	Years as TLP
Apache Kibble	18-10-2017	1
Apache Whimsical	20-05-2015	3
Apache Bahir	18-06-2016	2
Apache ORC	22-04-2015	4

NOTE: Years have been rounded down to previous full year as at 30th April 2019.

Chapter 4

Data Analysis

4.1 Apache HTTP Server Cultural Baseline

4.1.1 Indicator 1: Baseline Pony Factor Codebase

The following graph shows the Pony Factor for the Apache HTTP Server code base from 1996 until 2019.

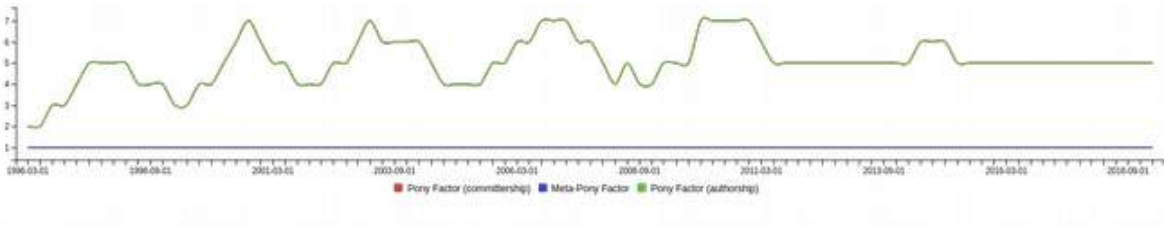

Figure 4.1: Apache HTTP Server Baseline Pony Factor Codebase 1996-2019

At the beginning of the project, in 1996 the Pony Factor is low (2). This is around the time of the creation of the Apache Group so it would be expected that the number of core maintainers would be low.

From 1996 onwards there is an increase in the Pony Factor, highlighting the growth in the number of people contributing to the project. Over time the Pony Factor rises and falls. The current figure for early 2019 is still high and remains constant.

The Pony Factor for committership not distinctly visible and follows the same profile as authorship.

The Meta Pony Factor remains constant at 1 because probably because all contributions are associated with the ASF affiliation via the committer email address.

4.1.2 Indicator 2: Baseline Pony Factor Email

The following graph shows the Pony Factor for the Apache HTTP Server mailing list from 1995 until 2019.

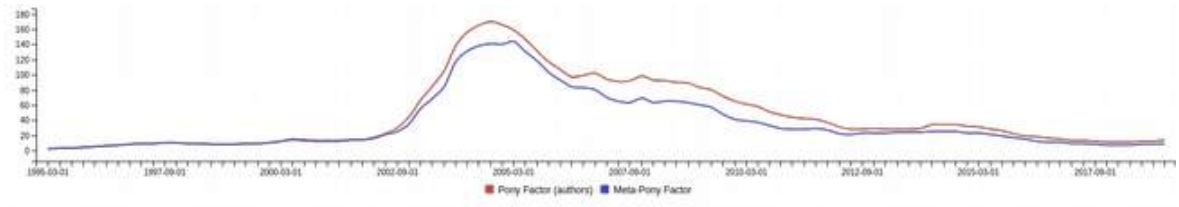

Figure 4.2: Apache HTTP Server Baseline Pony Factor Email 1996-2019

As with the codebase, the Pony Factor for the Apache HTTP Server mailing lists in 1995 is low (2) highlighting that there were a limited number of people active on the mailing lists. The number of active mailing list contributors gradually rises to a peak of up to 170 during 2005. It then reduces to 11 and is now rising again and is currently at 14.

The Meta Pony factor increases because email contributions are more flexible and affiliations can be more easily captured. It follows the same curve as the Pony Factor and has now reduced and stabilised at 9.

4.1.3 Indicator 3: Baseline Contributor Experience

The following graph shows a breakdown the length of time contributors have been contributing to the codebase:

The Apache HTTP Server project is over 20 years old and the oldest ASF project

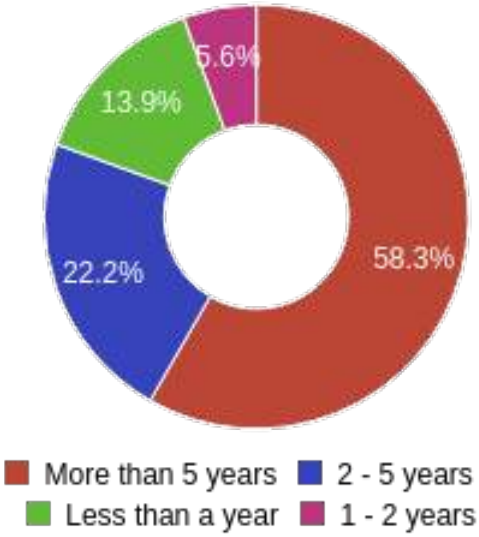

Figure 4.3: Apache HTTP Server Baseline Contributor Experience

- Over 58% of their contributors have been contributing to the project for more than 5 years
- Only 6% of their contributors have less than a year's experience.

4.1.4 Indicator 4: Baseline Contributor Retention Codebase

The following graph shows how many people have been retained as part of the community and contribute to the Apache HTTP Server codebase.

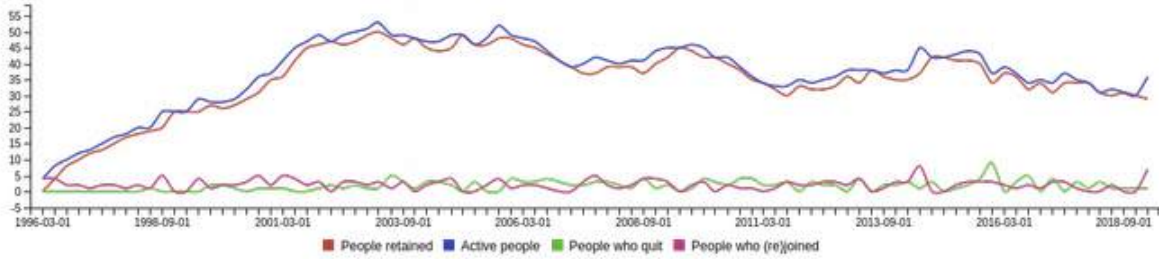

Figure 4.4: Apache HTTP Server Baseline Contributor Retention Codebase 1996-2019

The number of active people and those retained follow a similar curve. This is expected as those who are continually active are being retained by the community.

There is also appears to be a relationship between the people who have left and

those who rejoin as it has remained at a stable level across the life of the project.

As at 2019, both contributor retention and people rejoining have started to rise.

4.1.5 Indicator 5: Baseline Contributor Retention Email

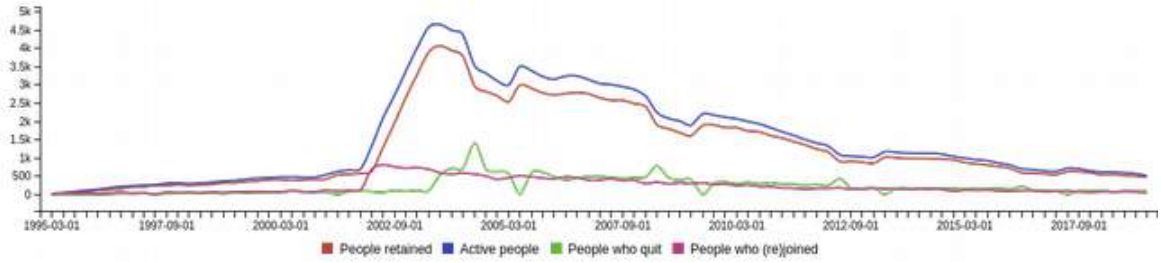

Figure 4.5: Apache HTTP Server Baseline Contributor Retention Email 1996-2019

During the beginning of the project the number of active people and those being retained was growing slowly until around 2001 when this substantially increased. It is also interesting to note during the main activity peak, there was also the highest number of people leaving the project. Both activity and retention have stabilised at a reduced level.

4.1.6 Indicator 6a: Baseline Sentiment Analysis as at 30/04/2019

The following graph shows the mood analysis of the Apache HTTP Server project on 30th April 2019.

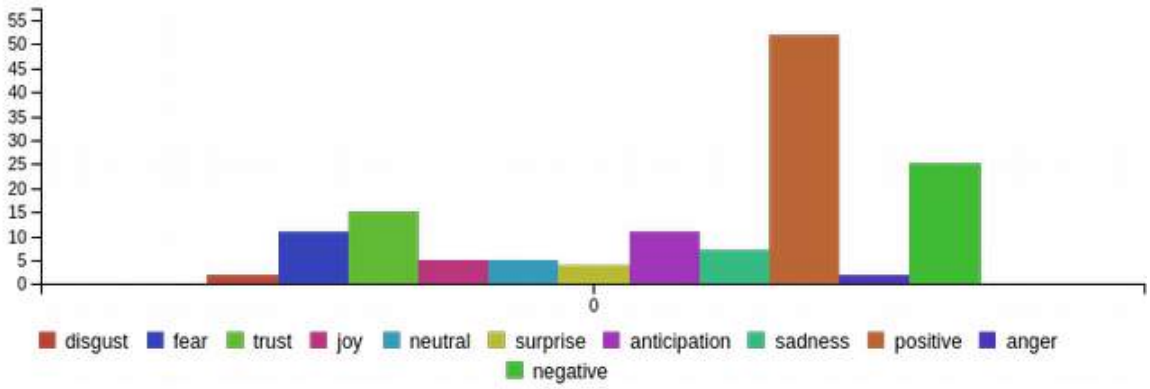

Figure 4.6: Apache HTTP Server Baseline Sentiment Analysis 30/04/2019

The highest mood sentiment is positivity (52) negativity (25) is second, followed by trust (15), fear (11) and anticipation (11).

The bottom 5 sentiments showing are neutral (5), joy (5), surprise (4), anger (2) and disgust (2)

4.1.7 Indicator 6b: Baseline Comparative Sentiment Analysis as at 30/04/2019

The following graphs show the comparative mood analysis of the Apache HTTP Server project as at 30th April 2019

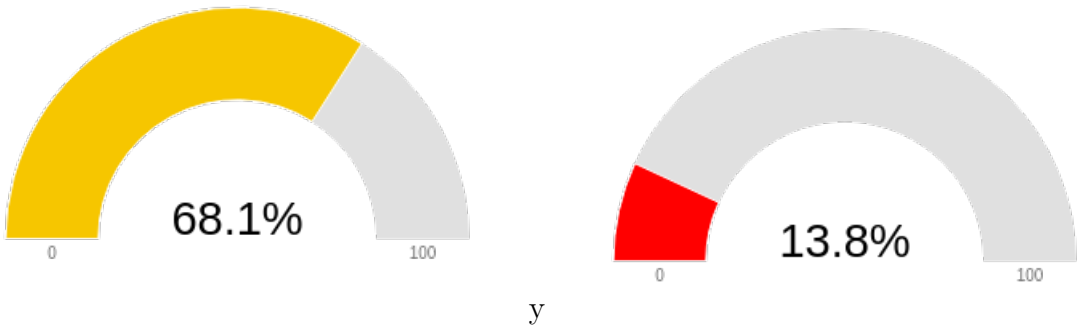

Figure 4.7: Apache HTTP Server vs. ASF Comparative Mood

The above graphs show the relative mood of Apache HTTP Server. The relative mood based on their communications is generally positive (68.1%)

When this mood is compared against the other ASF projects in Apache Kibble (currently 63) the mood showing (13.8%) indicative of the general level of mood intensity. So this means that the mood level of Apache HTTP Server is a lot lower in intensity than other ASF projects.

4.1.8 Indicator 7: Baseline Sentiment Analysis Over Time

The following graph shows the mood analysis of the Apache HTTP Server project from November 2018 – April 2019.

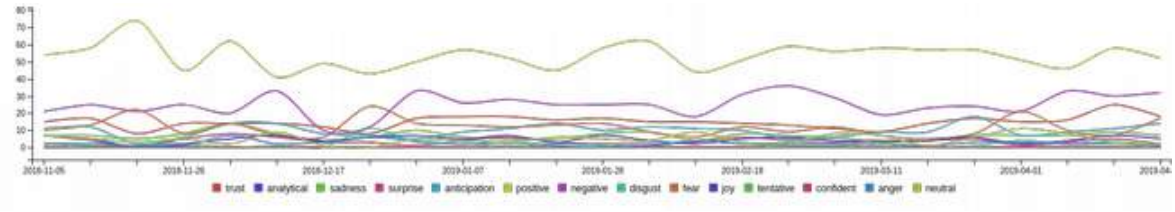

Figure 4.8: Apache HTTP Server Baseline Mood Analysis November 2018 - April 2019

The analysis over time shows that the positive communication style is established and although it rises and falls over time, it is the strongest mood expressed.

4.1.9 Indicator 8: Baseline Key Phrase Extraction

The following graph shows the key phrase extraction analysis of the most common phrases used in the Apache HTTP Server project as at 30th April 2019.

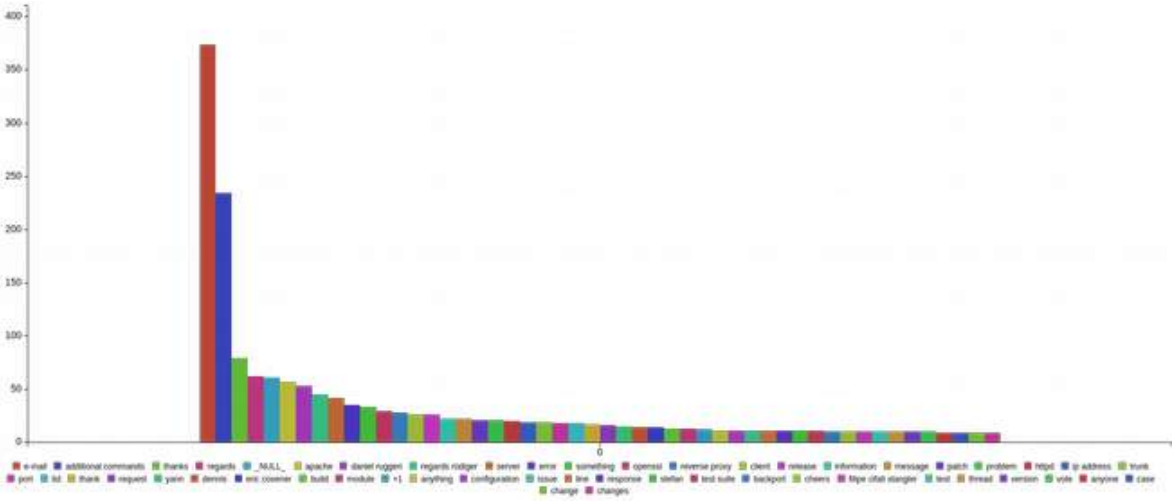

Figure 4.9: Apache HTTP Server Baseline Key Phrase Extraction: 30/04/2019

The phrases can be broken down into several areas:

- General or standard (e.g. email, thanks, regards, names etc.)
- Technical discussions (e.g. HTTPd, server, openssl, release, client, trunk, ip address, reverse proxy, _NULL_, virtual host)
- Collaborative (e.g. error, test suite, problem, thank, test, additional commands, something, request, backport)

- Cultural (+1, apache, vote, thread)

Some phrases appear to be aligned to ASF values as follows:

- Openness: Technical conversations in the open
- Consensus: +1 = indication of consensus
- Collaboration: Polite and collaborative communication

NOTE: The KPE appears to be in alignment with the mood analysis of positive communication. No negative phrases are coming out as significant.

4.2 Incubated non-Chinese Initiated Projects

4.2.1 Indicator 1: Incubated Non-Chinese Initiated Projects Pony Factor Codebase

The following graph shows the Pony Factor of the codebase for sub group of incubated non- Chinese initiated projects that have been through the Apache incubation process.

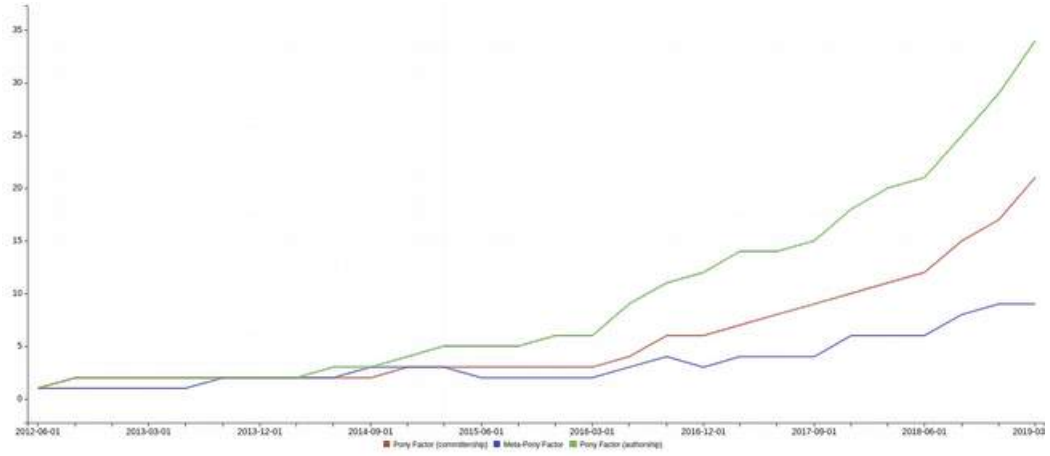

Figure 4.10: Incubated non-Chinese initiated Projects Pony Factor Codebase 2012 - 2019

As with Apache HTTP Server during the initial project stages the Pony Factor is low (1) and gradually increases over time to its current level (34). Notice that the actual shape of the graph is not the same shape as the Apache HTTP Server curve. This one is more gradual with no sharp changes.

The Pony Factor for committership follows the same trend as the curve for authorship except with a significant difference (21 instead of 34).

Notice that the Meta Pony Factor is not constant at 1 but is gradually increasing and is currently at 9. This is interesting because it is the first metric showing more than one affiliation for contributors to the codebase. This could be as result other version control repositories (e.g. github) being used for newer projects rather than Apache Subversion.

4.2.2 Indicator 2: Incubated Non-Chinese Initiated Projects Pony Factor Email

The following graph shows the Pony Factor for the incubated non-Chinese initiated projects.

NOTE: The mailing list statistics cover the period 2016 to 2019.

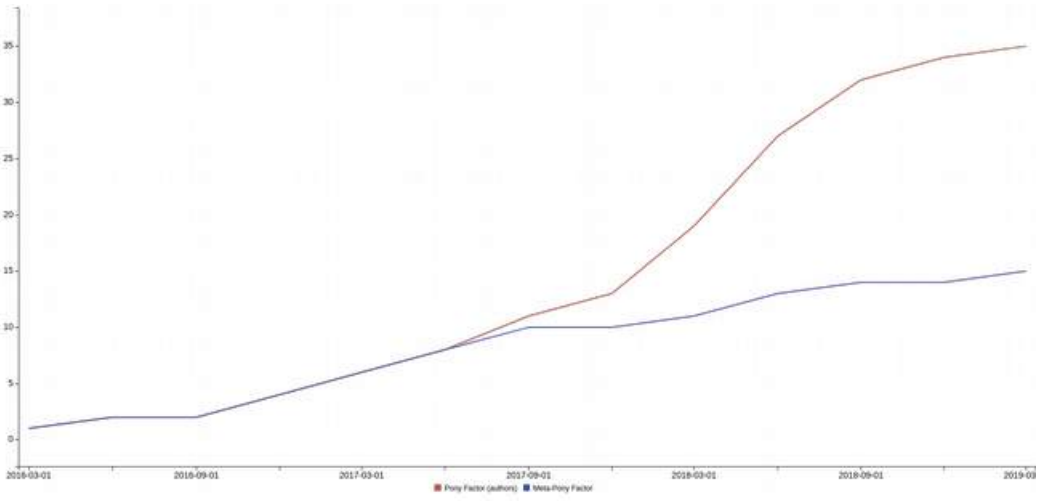

Figure 4.11: Incubated non-Chinese initiated Projects Pony Factor Email 2012 - 2019

As with the codebase, the mailing list Pony Factor for the incubated non-Chinese

initiated projects starts low (1). The number of active mailing list contributors gradually rises to its current peak (35).

This curve is unlike the Apache HTTP Server curve for this indicator. It is rising whereas Apache HTTP Server was reducing.

4.2.3 Indicator 3: Incubated Non-Chinese Initiated Projects Contributor Experience

The following graph shows a breakdown the length of time contributors have been contributing to the codebase:

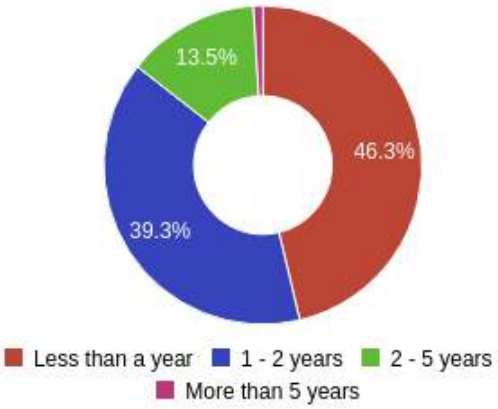

Figure 4.12: Incubated non-Chinese initiated Projects Contributor Experience

- Over 46% of their contributors have been contributing to the project for less than a year
- Over 39% of contributors have been contributing for 1 -2 years.
- Over 13% of contributors have been contributing for between 2 to 5 years.
- Less than 1% have more than 5 years experience.

NOTE: With over 85% of contributors with less than 2 years experience, this is a significant difference from the Apache HTTP Server profile.

4.2.4 Indicator 4: Incubated Non-Chinese Initiated Projects Contributor Retention Codebase

The following graph shows how many people have been retained as part of the community and contribute to the codebase.

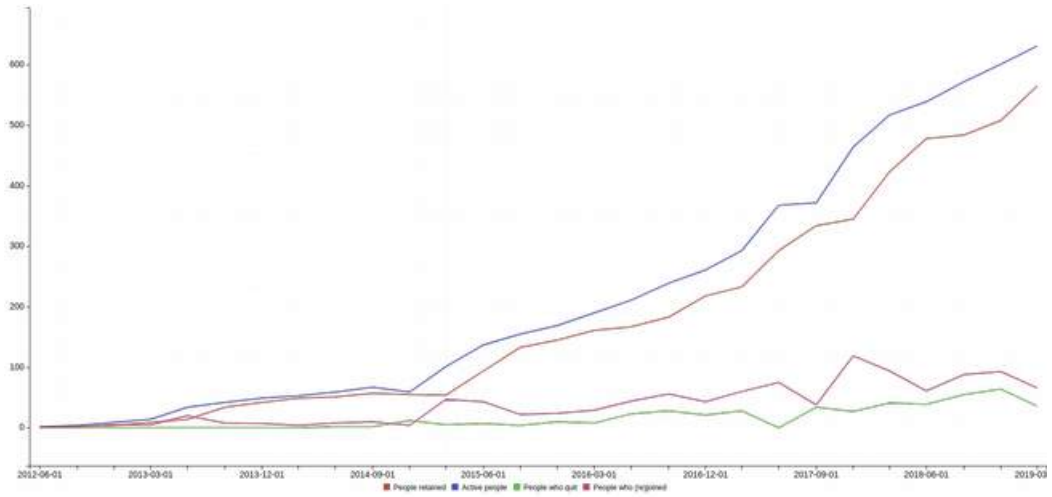

Figure 4.13: Incubated non-Chinese initiated Projects Contributor Retention Codebase 2012 - 2019

This shows a gradual increase of active people and those that are retained over time. With the retention rate still rising, it is very different profile from the Apache HTTP Server profile.

As at the time of writing, the current 2019 figures show:

- 631 active people
- 535 people retained
- 66 people rejoined
- 36 people quit

4.2.5 Indicator 5: Incubated Non-Chinese Initiated Projects Contributor Retention Email

The following graph shows how many people have been retained as part of the community and contribute to the mailing list discussions.

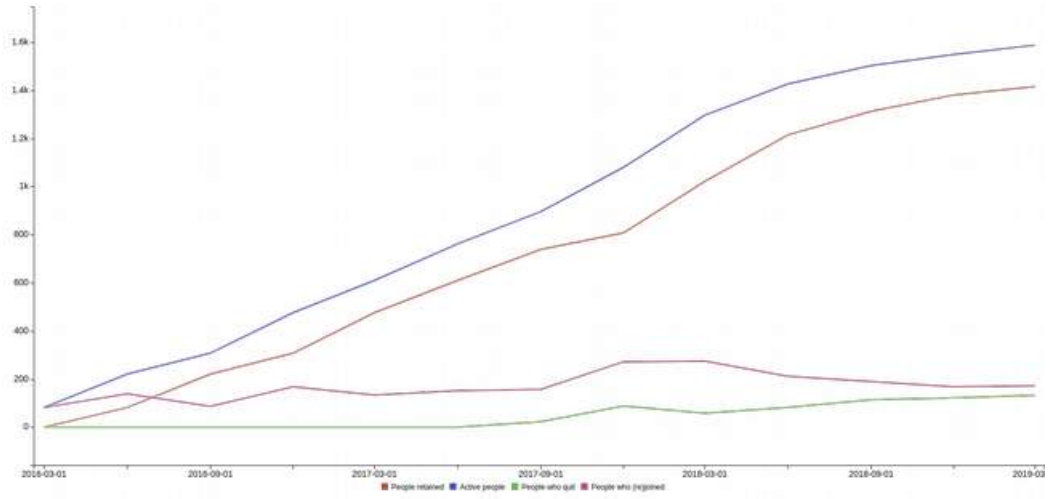

Figure 4.14: Incubated non-Chinese initiated Projects Contributor Retention Email 2012 - 2019

This graph does not look the Apache HTTP Server baseline. Instead it has a gradual increase over time that is still rising. This means that the number of contributors is actively growing and that they are being retained.

4.2.6 Indicator 6a: Incubated Non-Chinese Initiated Projects Sentiment Analysis as at 30/04/2019

The following graph shows the consolidated mood analysis of the Non Chinese initiated incubated projects as at 30th April 2019.

The highest mood sentiment is positivity (62) , negativity (15) is second, followed by trust (12) and anticipation (10) and joy (7).

The bottom five sentiments showing are sadness (5), surprise (3), neutral (3), anger (2) and disgust (1).

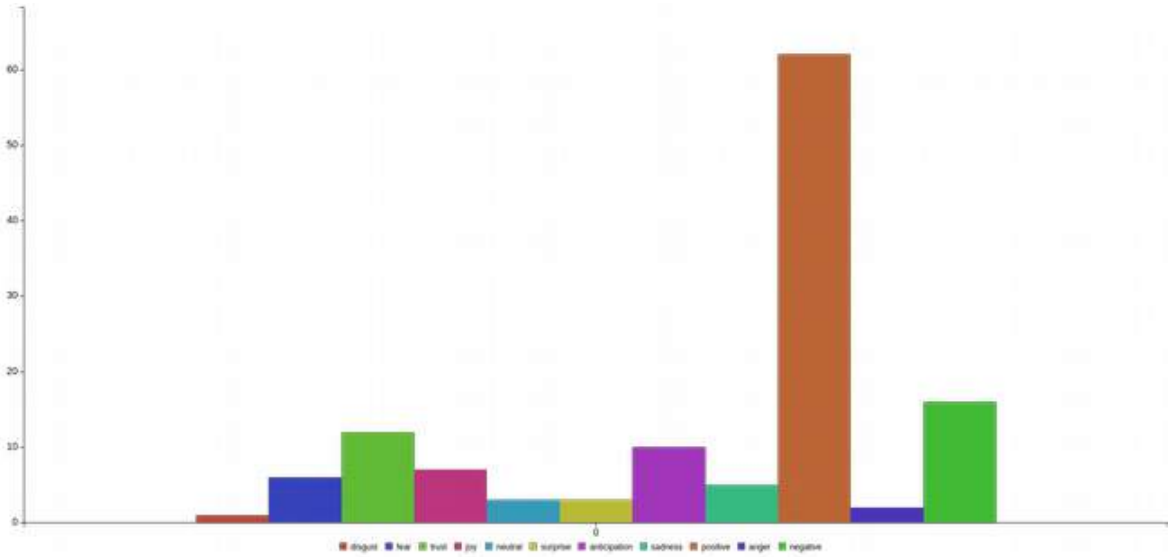

Figure 4.15: Incubated non-Chinese initiated Projects Sentiment Analysis 30/04/2019

NOTE: The 3 strongest sentiments are the same as for Apache HTTP Server and they are similarly distributed. There are some fluctuations in the other sentiments but they are not significant.

4.2.7 Indicator 6b: Incubated Non-Chinese Initiated Projects Comparative Sentiment Analysis as at 30/04/2019

The following graphs show the comparative mood analysis of the Incubated non-Chinese initiated projects as at 30th April 2019.

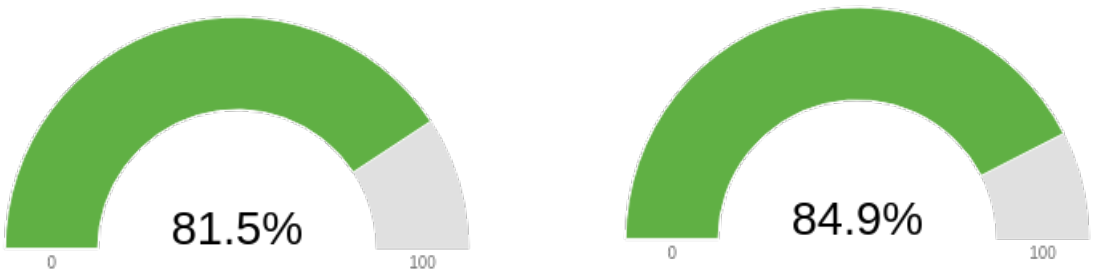

Figure 4.16: Incubated non-Chinese initiated Projects vs. ASF Comparative Mood

The above graphs show the relative mood of the incubated non-Chinese initiated projects. The relative mood based on their communications is very positive (81.5)

When this mood is compared against other ASF projects in Apache Kibble (currently 63) the mood (84.9) is showing as indicative of the general level of ASF project mood intensity. So this means that the incubated non-Chinese initiated projects exhibit a similar level of mood intensity as other ASF projects.

4.2.8 Indicator 7: Incubated non-Chinese Initiated Projects Sentiment Analysis Over Time

The following graph shows the mood analysis of the incubated non-Chinese initiated projects from November 2018 – April 2019.

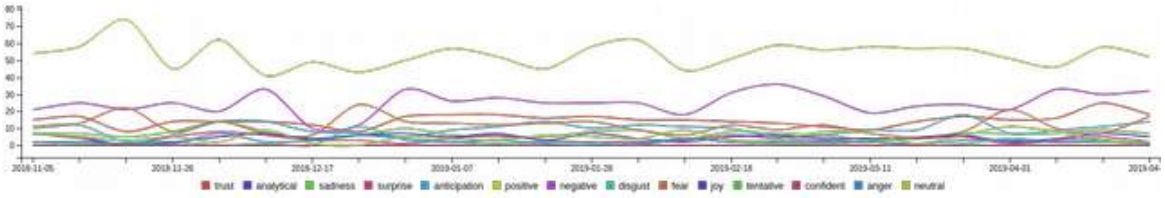

Figure 4.17: Incubated non-Chinese initiated Projects Sentiment Analysis Over Time November 2018 - April 2019

The analysis over time shows that the positive communication style is established and is the strongest mood expressed. This is very similar to the Apache HTTP Server baseline model.

4.2.9 Indicator 8: Incubated non-Chinese Initiated Projects Key Phrase Extraction

The following graph shows the key phrase extraction analysis of the most common phrases used in the incubated non-Chinese initiated projects as at 30th April 2019.

The phrases can be broken down into several areas:

- General or standard (e.g. message, thanks, email, regards, names etc.)
- Technical discussions (e.g. _NULL_, additional commands, release, error, file, repository. code, tests)

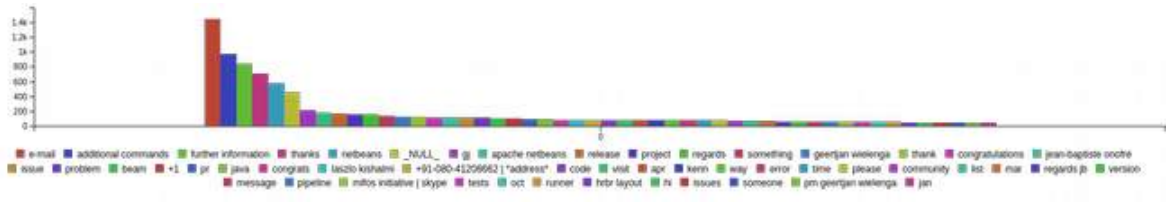

Figure 4.18: Incubated non-Chinese initiated Projects Key Phrase Extraction: 30/04/2019

- Collaborative (e.g. additional comments, further information, congratulations, project, documentation, please, attachment, improvement reporter)

- Cultural (e.g. community, +1, list)

Several cultural indicators are aligned to ASF values as follows:

- Openness: Technical conversations in the open.

- Consensus: +1 = indication of consensus.

- Collaboration: Polite communication, information requests, issue assignment, problem resolution.

- Community: mailing list is the communication medium.

This KPE profile is very similar to Apache HTTP Server and many common phrases are being used including 'community'. The +1 consensus indicator is culturally significant.

NOTE: The KPE is also in alignment with the mood analysis which highlights a positive communication style. There are a lot of collaborative phrases that are being extracted and no negative phrases are coming out as significant.

4.3 Incubated Chinese Initiated Projects

4.3.1 Indicator 1: Incubated Chinese Initiated Projects Pony Factor Codebase

The following graph shows the Pony Factor of the codebase for sub group of incubated Chinese initiated projects that have been through the Apache incubation process.

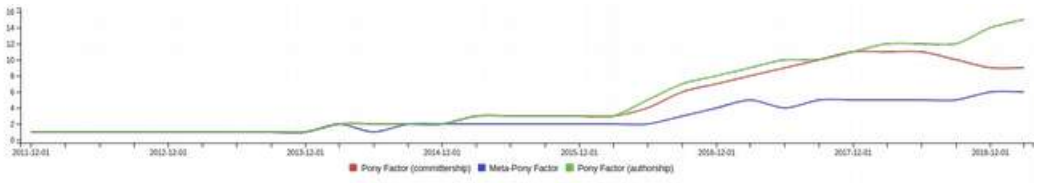

Figure 4.19: Incubated Chinese initiated Projects Pony Factor Codebase 2011 - 2019

As with Apache HTTP Server during the initial project stages the Pony Factor is low (1) and gradually increases over time to its current level (15) which is half that of the incubated non-Chinese initiated projects. Notice that the actual shape of the graph is not the same as the Apache HTTP Server curve but it does resemble the general curve of the incubated non-Chinese initiated projects.

The Pony Factor for committership follows the same trend as the curve for authorship and matches it closely before dropping off to the current level of 9.

Notice that the Meta Pony Factor is not constant at 1 but is gradually increasing and is currently at 6.

4.3.2 Indicator 2: Incubated Chinese Initiated Projects Pony Factor Email

The following graph shows the Pony Factor for the incubated Chinese initiated projects.

NOTE: The mailing list statistics cover the period 2015 to 2019.

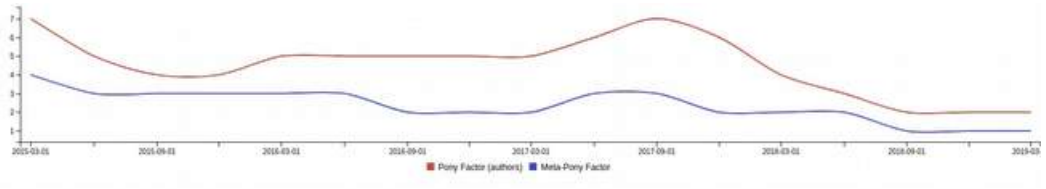

Figure 4.20: Incubated Chinese initiated Projects Pony Factor Email 2015 - 2019

This is looks very different to both Apache HTTP Server and the incubated non-Chinese initiated projects. The Pony Factor starts at its highest and then reduces, peaks and has now dropped off to a low of 2.

This highlights that the main email traffic is being generated by very few people. This is interesting because it could indicate that conversations are perhaps happening elsewhere (i.e. not all in the community are fluent in English) or that perhaps the community needs very little communication to initiate any work.

4.3.3 Indicator 3: Incubated Chinese Initiated Projects Contributor Experience

The following graph shows a breakdown the length of time contributors have been contributing to the codebase:

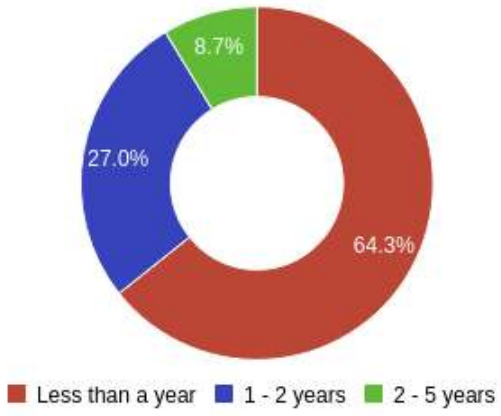

Figure 4.21: Incubated Chinese initiated Projects Contributor Experience

- Over 64% of contributors have been contributing to the project for less than a year

- 27% of contributors have been contributing for between 1 - 2 years
- Over 8% of contributors have 2 - 5 years experience.

NOTE: This is a significant difference from Apache HTTP Server profile. The new contributor rate is even higher than the incubated non-Chinese initiated projects.

4.3.4 Indicator 4: Incubated Chinese Initiated Projects Contributor Retention Codebase

The following graph shows how many people have been retained as part of the community and contribute to the codebase.

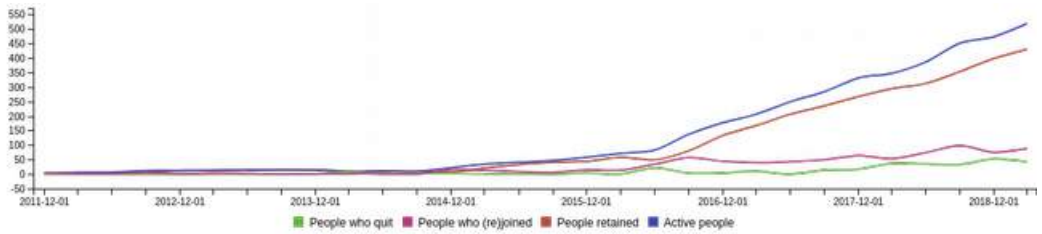

Figure 4.22: Incubated Chinese initiated Projects Contributor Retention Codebase 2011 - 2019

This shows a flat curve with before the projects entered Apache Incubator in 2014. After which there is a gradual increase of active people and those retained over time. With the retention rate still rising, it is very different profile from Apache HTTP Server profile. As at the time of writing, the current 2019 figures show:

- 518 active people
- 430 people retained
- 88 people rejoined
- 43 people quit

NOTE: The code repository was created in 2011 but the projects entered incubator from 2014 onwards and this is where the contributor retention begins. This

seems to indicate that something happened to the community as a result of entering and remaining in incubation.

4.3.5 Indicator 5: Incubated Chinese Initiated Projects Contributor Retention Email

The following graph shows how many people have been retained as part of the community and contribute to the mailing list discussions.

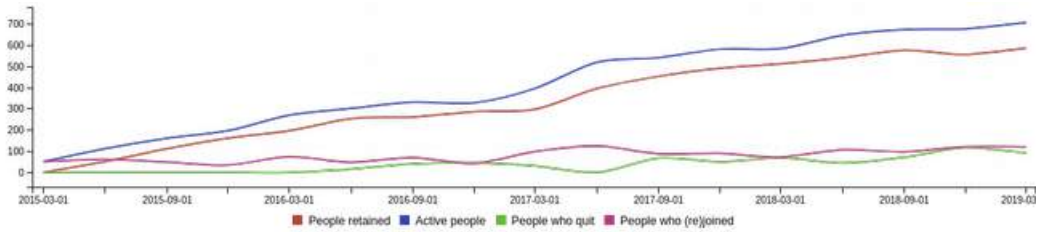

Figure 4.23: Incubated Chinese initiated Projects Contributor Retention Email 2012 - 2019

This graph does not look the Apache HTTP Server baseline but is similar to the Incubator control group of non-Chinese initiated projects. It increases gradually over time and is still rising. This means that the number of contributors is actively growing and that they are being retained.

NOTE: The number of contributors is is 705 which approximately half of those in the Apache Incubator control group, 1588.

4.3.6 Indicator 6a: Incubated Chinese Initiated Projects Sentiment Analysis as at 30/04/2019

The following graph shows the consolidated mood analysis of the Non Chinese initiated incubated projects as at 30th April 2019.

The highest mood sentiment is positivity (64), second is trust (15), followed by negativity (12), and anticipation (11) and fear (9).

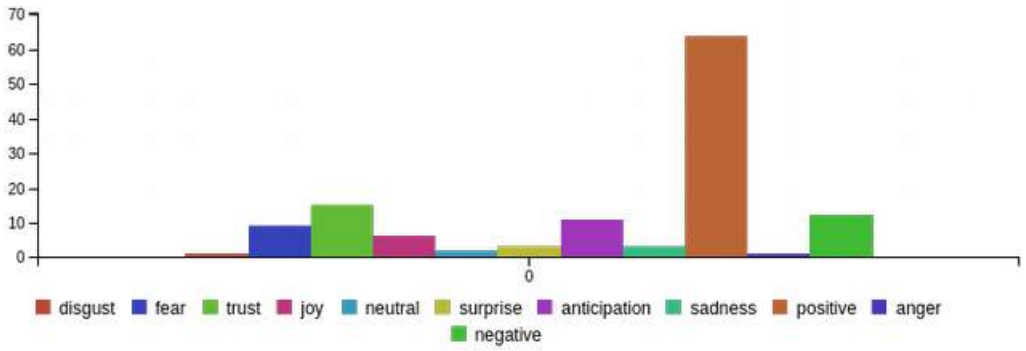

Figure 4.24: Incubated Chinese initiated Projects Sentiment Analysis 30/04/2019

The bottom five sentiments showing are sadness (3), surprise (3), neutral (2), anger (1) and disgust (1).

NOTE: The 3 strongest sentiments are the same as for Apache HTTP Server and the incubated non-Chinese initiated projects and are similarly distributed. There are some fluctuations in the other sentiments but they are not significant.

4.3.7 Indicator 6b: Incubated Chinese Initiated Projects Comparative Sentiment Analysis as at 30/04/2019

The following graphs show the comparative mood analysis of the Incubated Chinese initiated projects as at 30th April 2019.

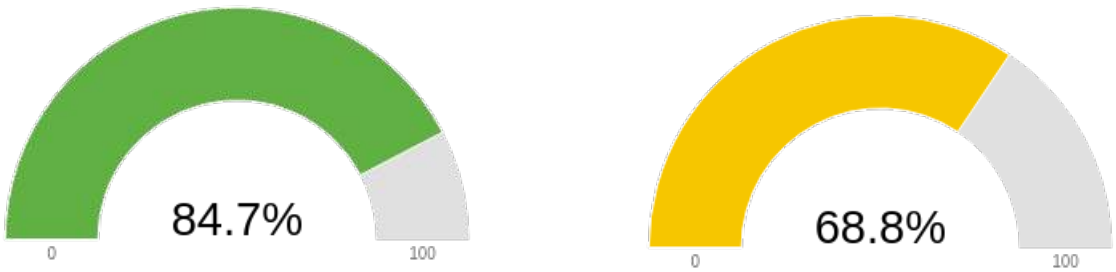

Figure 4.25: Incubated Chinese initiated Projects vs. ASF Comparative Mood

The above graphs show the relative mood of the incubated Chinese initiated projects. The relative mood based on their communications is very positive (84.7) and is higher than that of the incubated non-Chinese initiated projects.

When this mood is compared against the other ASF projects in Apache Kibble (currently 63) the mood (68.8) is showing that the mood being expressed is not at the same level of intensity as the other ASF projects. So this means that the incubated Chinese initiated projects are not as intense in expressing their moods as the other ASF projects.

4.3.8 Indicator 7: Incubated Chinese Initiated Projects Sentiment Analysis Over Time

The following graph shows the mood analysis of the incubated non-Chinese initiated projects from November 2018 – April 2019.

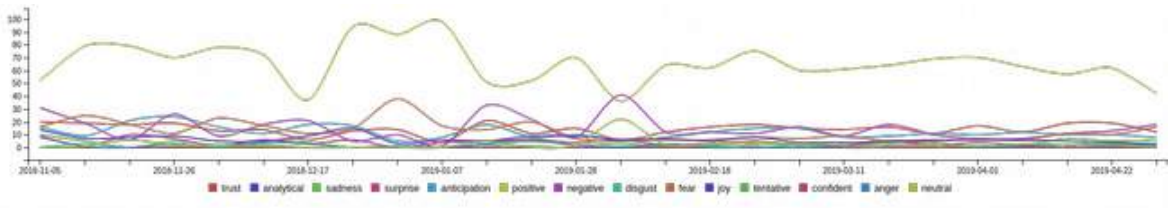

Figure 4.26: Incubated Chinese initiated Projects Sentiment Analysis Over Time November 2018 - April 2019

The analysis over time shows that the positive communication style is established and is the strongest mood expressed. This mood does tend to vary a lot more than Apache HTTP Server and the incubated non-Chinese initiated projects. On 4th February 2019, the negative mood was higher than the positive mood.

4.3.9 Indicator 8: Incubated Chinese Initiated Projects Key Phrase Extraction

The following graph shows the key phrase extraction analysis of the most common phrases used in the incubated Chinese initiated projects as at 30th April 2019.

The phrases can be broken down into several areas:

- General or standard (e.g. message, thanks, email, regards, names etc.)

Figure 4.27: Incubated Chinese initiated Projects Key Phrase Extraction: 30/04/2019

- Technical discussions (e.g. _NULL_, cube, error, measure, local exception, rocketbot, code, improvement reporter)
- Collaborative (e.g. use, project, discuss graduate, problem, issue, call, improvement reporter, contributor)
- Cultural (e.g. community, +1, list, apache)

Several cultural indicators are aligned to ASF values as follows:

- Openness: Technical conversations in the open
- Consensus: +1 = indication of consensus
- Collaboration: Polite communication, information requests, issue assignment, problem resolution
- Community: mailing list is the communication medium

4.4 Non-Chinese Projects Bypassing Incubation

4.4.1 Indicator 1: Non-Incubated non-Chinese Initiated Projects Pony Factor Codebase

The following graph shows the Pony Factor of the codebase for the four non-Ćhinese initiated projects that bypassed the Apache incubation process.

This graph does not resemble the any of the results from Apache HTTP Server baseline, the incubated non-Chinese initiated projects or the incubated Chinese initiated projects.

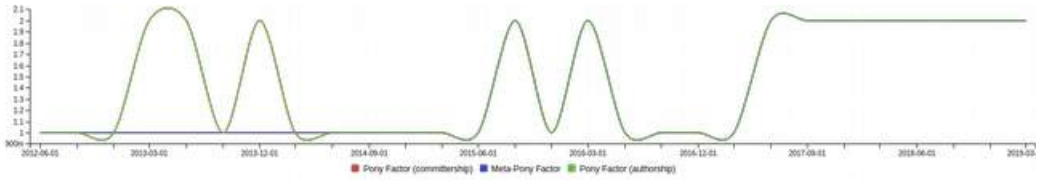

Figure 4.28: Non-Incubated non-Chinese initiated Projects Pony Factor Codebase 2012 - 2019

There seems to be a inconsistent increase and decrease in Pony Factor perhaps caused by key people being intermittently active. Another cause could be that contributors are not being retained.

4.4.2 Indicator 2: Non-Incubated non-Chinese Initiated Projects Pony Factor Email

The following graph shows the Pony Factor for the non-Chinese initiated projects that bypassed incubation.

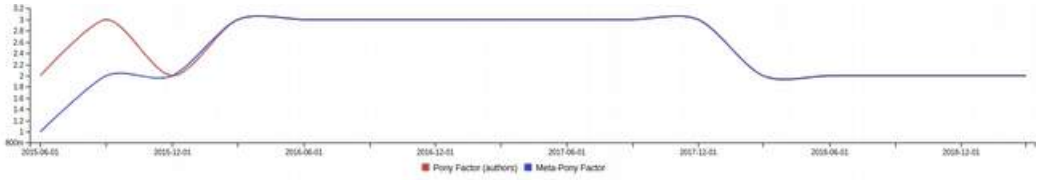

Figure 4.29: Non-Incubated non-Chinese initiated Projects Pony Factor Email 2015 - 2019

The curve for this indicator looks very different from the incubated non-Chinese initiated projects and incubated Chinese initiated projects. It does however have some similar characteristics to the Apache HTTP Server profile where the Pony Factor rises, flattens, then reduces to stabilise at 2.

4.4.3 Indicator 3: Non-Incubated non-Chinese Initiated Projects Contributor Experience

The following graph shows a breakdown the length of time contributors have been contributing to the codebase:

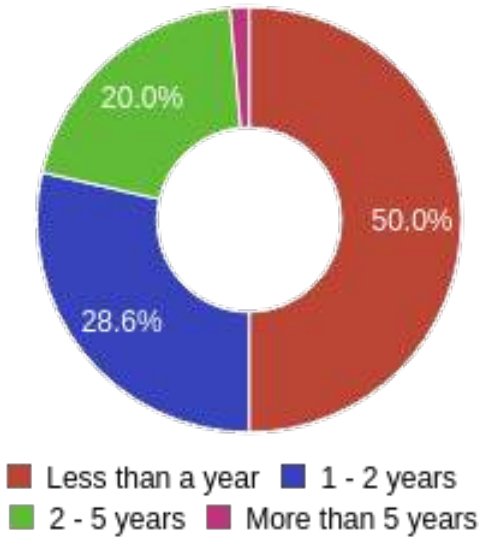

Figure 4.30: Non-Incubated non-Chinese initiated Projects Contributor Experience

- 50% of contributors have been contributing to the project for less than a year
- Over 28% of contributors have been contributing for between 1 - 2 years
- Over 20% of contributors have 2 - 5 years experience.
- Less than 2% have been involved for over 5 years

There seems to be a good mix and flow of contributors with the majority of them being less than a year.

4.4.4 Indicator 4: Non-Incubated non-Chinese Initiated Projects Contributor Retention Codebase

The following graph shows how many people have been retained as part of the community and contribute to the codebase.

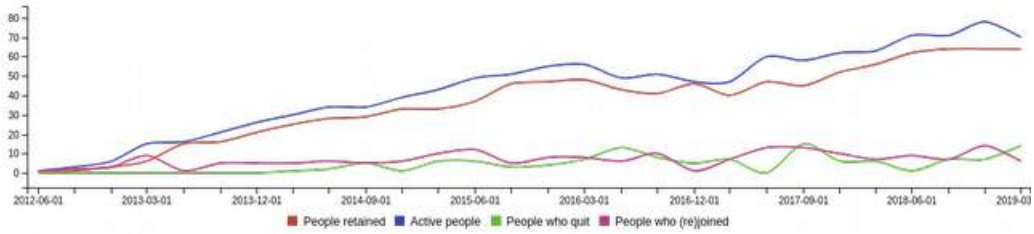

Figure 4.31: Non-Incubated non-Chinese initiated Projects Contributor Retention Codebase 2011 - 2019

This shows a gradual increase with slight dips. It is a very similar profile to the incubated non-Chinese initated projects and the incubated Chinese initiated projects.

As at the time of writing, the current 2019 figures show:

- 70 active people
- 64 people retained
- 6 people rejoined
- 14 people quit

NOTE: These projects have a lot smaller communities compared with the incubated non-Chinese initiated projects (1500 approx) and the incubated Chinese initiated projects (700 approx).

4.4.5 Indicator 5: Non-Incubated non-Chinese Initiated Projects Contributor Retention Email

The following graph shows how many people have been retained as part of the community and contribute to the mailing list discussions.

This graph does not look the Apache HTTP Server baseline but is similar to the incubated non-Chinese initiated projects and the incubated Chinese initiated projects. It increases gradually over time and is still rising. This means that the number of contributors is actively growing and that they are being retained.

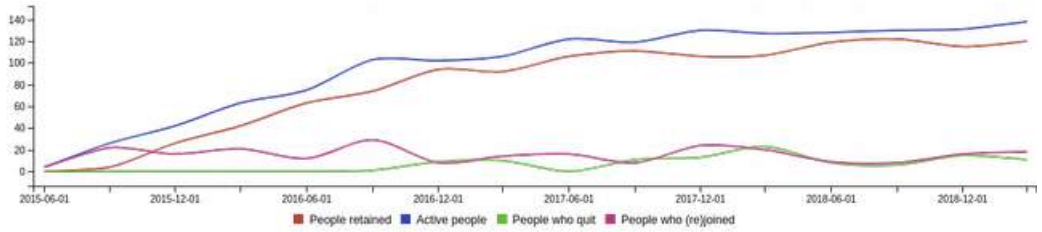

Figure 4.32: Non-Incubated non-Chinese initiated Projects Contributor Retention Email 2012 - 2019

4.4.6 Indicator 6a: Non-Incubated non-Chinese Initiated Projects Sentiment Analysis as at 30/04/2019

The following graph shows the consolidated mood analysis of the non-Chinese initiated projects that bypassed incubation as at 30th April 2019.

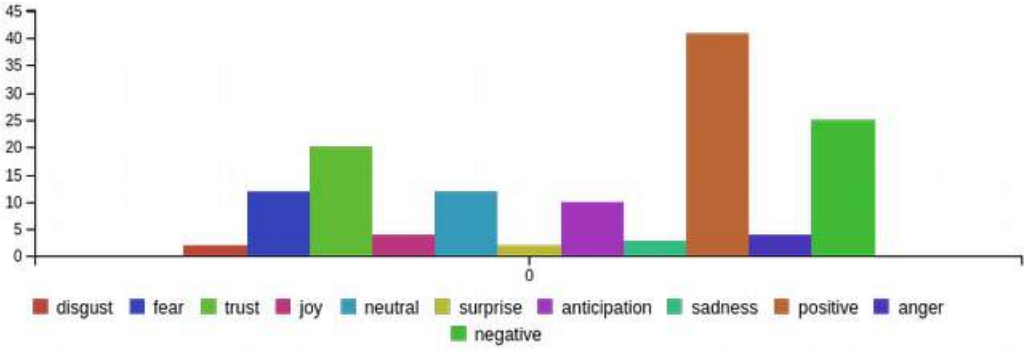

Figure 4.33: Non-Incubated non-Chinese initiated Projects Sentiment Analysis 30/04/2019

The highest mood sentiment is positivity (45), second is negativity (25), followed by trust (20), and neutral (21) and fear (12).

The bottom five sentiments showing are joy (4), anger (4), sadness (3), surprise (2) and disgust (2).

4.4.7 Indicator 6b: Non-Incubated non-Chinese Initiated Projects Comparative Sentiment Analysis as at 30/04/2019

The following graphs show the comparative mood analysis of the non-Chinese initiated projects that bypassed incubation as at 30th April 2019.

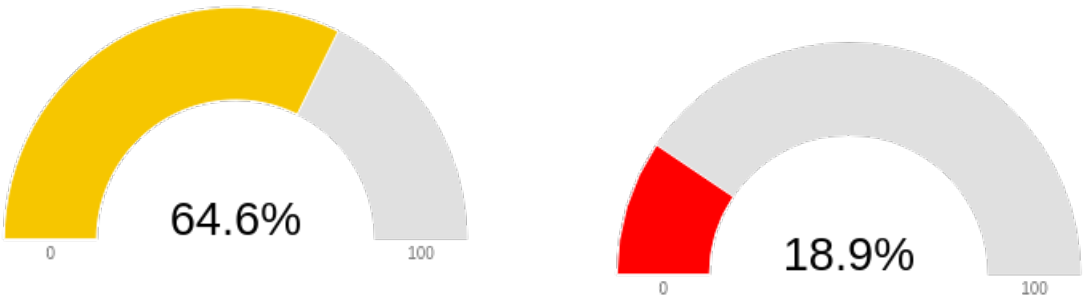

Figure 4.34: Non-Incubated non-Chinese initiated Projects vs. ASF Comparative Mood

The above graphs show the relative mood of the non-Chinese initiated projects that bypassed incubation. The relative mood based on their communications is very mixed (64.6) and is lower than both groups of incubated projects, Chinese initiated and non-Chinese initiated.

When this mood is compared to the other ASF projects in Apache Kibble (currently 63) the mood (18.9) shows that the mood being expressed is not at the same level of intensity as the other ASF projects. The sentiments here are a lot weaker.

4.4.8 Indicator 7: Non-Incubated non-Chinese Initiated Projects Sentiment Analysis Over Time

The following graph shows the mood analysis of the non-Chinese Initiated projects that bypassed incubation from November 2018 – April 2019.

The mood over time looks erratic. The most dominant mood being displayed is positivity. There are times where positivity is low and yet there is only one distinct peak, for a short period when negativity is the most dominant.

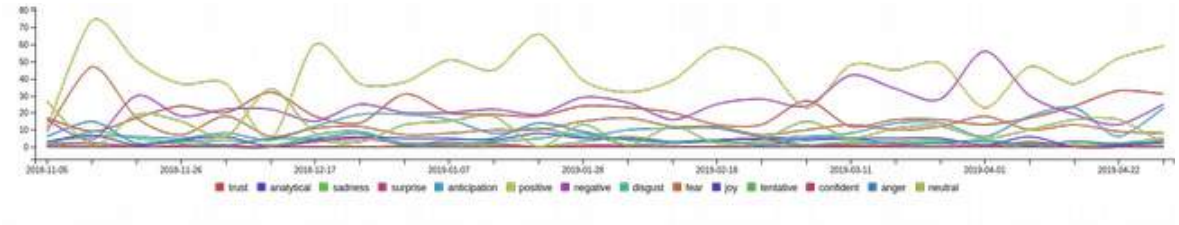

Figure 4.35: Non-Incubated non-Chinese initiated Projects Sentiment Analysis Over Time November 2018 - April 2019

4.4.9 Indicator 8: Non-Incubated non-Chinese Initiated Projects Key Phrase Extraction

The following graph shows the key phrase extraction analysis of the most common phrases used in the non-Chinese initiated projects that bypassed incubation as at 30th April 2019.

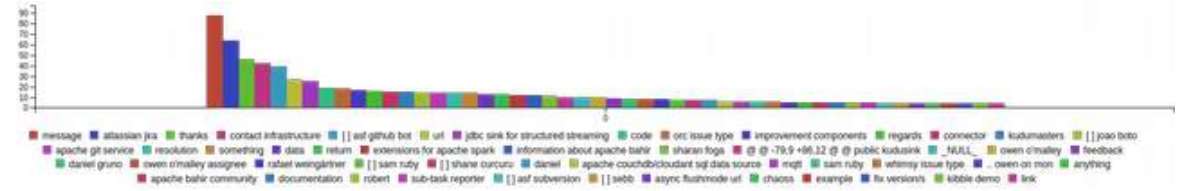

Figure 4.36: Non-Incubated non-Chinese initiated Projects Key Phrase Extraction: 30/04/2019

The phrases can be broken down into the following areas:

- General or standard (e.g. message, thanks. names etc.)
- Technical discussions (e.g. contact infrastructure, code, connector, jdbc for structured streaming, apache git service data source)
- Collaborative (e.g. url, issue type, improvement components, resolution, data, feedback, documentation, feedback)
- Cultural (apache bahir community)

Several cultural indicators aligned to ASF values are as follows:

- Openness: Technical conversations in the open

- Collaboration: Polite communication, information requests, issue assignment, problem resolution
- Community: apache bahir community

Very little cultural language is being used and most significantly there is no +1 consensus indicator is coming out in KPE analysis for these projects.

Chapter 5

Results

The purpose of this study was to gain a better understanding of the open source culture at Apache Software Foundation and analyse a project that has a different cultural profile to see if cultural changes occur as part of the project evolution and graduation to Top Level Project.

This study focuses on four Chinese contributed projects, all of which have undergone or are currently undergoing incubation at the Apache Software Foundation.

Twelve ASF projects were selected and divided into three groups for comparison against the Apache HTTP Server cultural baseline. They were also compared against each other. The groups were as follows:

- Group 1: Four Apache projects which have undergone incubation that have been initiated by Chinese contributors

- Group 2: Four Apache projects which have undergone incubation that have not been initiated by Chinese contributors

- Group 3: Four Apache projects which did not undergo incubation that have not been initiated by Chinese contributors

5.1 Apache HTTP Server Cultural Baseline

The following results below show the comparison of each group against the Apache HTTP Server baseline.

5.1.1 Pony Factors

None of the three groups appeared to follow the Apache HTTP Server baseline for the codebase or email. There was a general difference between codebase authorship and committership, except for the non incubated projects, showing that they are perhaps not as frequent in recognising merit based on contributor activity.

Where there was a difference the Chinese initiated projects showed the lowest gap between codebase authorship and committership perhaps indicating that they are more generous in recognising merit. This could potentially be a manifestation of collectivist group culture where people joining the group are more readily recognised.

The Pony Factor for email was different across all three groups with none of them matching the Apache HTTP Server baseline for this indicator.

5.1.2 Contributor Retention

None of the groups appear to follow the Apache HTTP Server baseline in this area. The majority of contributors for all three groups were made up of newer contributors. In the Chinese initiated projects, over 90% of contributors have been contributing for less than 2 years compared to 85% in the incubated non-Chinese initiated projects, 70% in the non-Incubated projects and 19% for Apache HTTP Server.

This means that all three groups are successful at attracting people to contribute and have a good flow of people coming into their communities.

Contributor retention is also very high which is significantly different to the Apache HTTP baseline.

5.1.3 Sentiment Analysis

All three groups of projects have similar profiles to the Apache HTTP Server baseline. The communication style was generally positive and although negativity was listed in the top 5 for all groups, it did not dominate the projects in the long term.

Over time the Chinese initiated projects and the non-Incubated projects moods were not as consistent as Apache HTTP Server.

Looking at the comparative moods and the intensity of the moods, only one of the groups followed a similar comparative sentiment to Apache HTTP Server and this was the non incubated projects. This shows that the sentiments for Apache HTTP Server and the non Incubated projects were a lot lower than all the other projects.

5.1.4 Key Phrase Analysis

All three groups of projects appeared to have similar profiles to the Apache HTTP Server baseline. They all showed elements of standard everyday communication and technical interactions. It was interesting to see that they also showed indications of cultural expression of ASF values, such as openness, collaboration and community.

The most significant cultural element which appeared in the baseline as well as two out of the three groups was the "+1" indicator. This is unique to the ASF as an indicator of consensus.

The only group where the "+1" indicator did not appear was the non-incubated project group.

5.2 Incubated Chinese vs. Incubated non-Chinese

The following results show the comparison of the incubated Chinese initiated projects against the the incubated non-Chinese projects.

5.2.1 Pony Factors

This Pony Factor for the codebase of both incubated Chinese and incubated non-Chinese projects look similar. They are both increasing and the curve of the non-Chinese initiated projects is steeper, showing faster growth.

In the Chinese initiated projects the gap between codebase authorship and committership over time is a lot lower than the non-Chinese initiated projects. For the Chinese projects, this could potentially be a manifestation of collectivist group culture where they are more generous in recognising merit.

Currently there is a gap between of authorship and committership. For the Chinese initiated projects is the difference is 6, and for the non-Chinese projects it is 13.

However looking at the email Pony Factor, the non-Chinese initiated projects have grown quickly and have stabilised at 35. The Pony Factor for the Chinese initiated projects has been decreasing and is currently stablised at 2. For the Chinese initiated projects, this means that the email messages on their combined mailing lists are generally centred around 2 people.

With one Pony Factor showing activity and community growth, and the other showing limited communication, for the Chinese initiated projects it could mean that the main email traffic is being generated by very few people. This is interesting because it could be an indication that conversations are perhaps happening elsewhere since not all the community are fluent in English, or that these communities need very little communication to initiate any work.

5.2.2 Contributor Retention

Both the Chinese initiated and non-Chinese initiated projects were good at attracting and retaining contributors.

In the case of the non Chinese initiated group, over 85% of their contributors

have less that 2 years experience. In the Chinese initiated projects, this figure is even higher, where 91% of their contributors have less than 2 years experience. This means the Chinese initiated projects are the most successful in attracting new contributors.

In both groups there are also people who have remained within the community and are currently within the 2 – 5 year range. As the projects mature, we would expect to see some growth in this figure as the new contributors age with the project.

Looking at the contributor retention over time, it can be clearly seen that, for the Chinese initiated projects, their contributor retention increased when they joined Apache Incubator. The contributor group size between the Chinese initiated and non-Chinese initiated is very similar:

- 631 active for incubated non-Chinese initiated projects
- 518 active for incubated Chinese initiated projects

For email, the number of contributors for the Chinese initiated projects is 705 which is less than half of the contributor number (1588) for the non-Chinese initiated projects.

5.2.3 Sentiment Analysis

The top mood for both the Chinese initiated and the non Chinese initiated groups was positivity. Both trust and negativity were also present as strong sentiments.

The analysis over time for both groups shows that the positive communication style is established and is the strongest mood expressed. For the Chinese initiated group the intensity of the positivity did vary a lot more than the non-Chinese projects and at one point during February 2019, the negative mood was higher than the positive one.

The overall mood for both groups was over 80% but when compared with other Apache projects the intensity of the sentiment was very different. The non-Chinese

initiated projects had a comparative mood of over 84% which means that their sentiments are very close to the ones in other ASF projects. For the Chinese initiated projects the comparative sentiment was a lot lower at 51%. This means that the Chinese initiated projects are not as intense in expressing their moods as other ASF projects.

In the case of the Chinese projects it could simply be cultural reticence that is generally encountered in using a foreign language. The type of phrasing and sentences used could perhaps lack emotion because they are focussed on communicating a specific technical tasks rather than social pleasantries.

5.2.4 Key Phrase Analysis

Both the Chinese initiated and the non-Chinese initiated groups have similar profiles. They showed elements of standard everyday communication and technical interactions. It was interesting to see that they also showed indications of cultural expression of ASF values such as openness, collaboration and community.

The most significant cultural element which appeared in both the Chinese and the non Chinese projects was the "+1" indicator. This is unique to the ASF as an indicator of consensus.

5.3 Non Incubated non-Chinese vs Incubated Chinese

5.3.1 Pony Factors

The Pony Factor for the codebase for the non-incubated projects shows an erratic pattern with inconsistent increases and decreases which could be potentially linked to key people being intermittently active. The Pony Factors for the Chinese initiated projects are increasing and still rising.

The non-incubated projects showed no difference between the Pony Factor for authorship and committership. This is probably because the people involved with the project are already experienced committers or long time mentors within the ASF. A key reason stated for bypassing incubation is that the community or the majority of contributors are already familiar with ASF culture.

The Chinese initiated projects show little or no gap over time, between codebase authorship and committership. This comparison is interesting because in the Apache HTTP Server baseline there was no difference between authorship and committership, and in the non-Incubated projects we see the same pattern. This appears to indicate that the Chinese initiated projects are more similar to the Apache HTTP Server and the non Incubated projects than to the non-Chinese initiated projects.

The Pony Factor for email for the non-Incubated projects rose, stabilised but has now reduced to be the same as the Chinese initiated projects (2) . This indicates that communication is centred around a small number of people.

5.3.2 Contributor Retention

There seems to be a good mix and flow of contributors in both the non-incubated and the Chinese initiated groups, with the majority of them being involved for less than a year.

With over 90% of contributors, the Chinese initiated projects are more successful than the non-incubated projects at attracting new contributors. The non-incubated projects have more of mix of contributors indicating that contributors transition to staying involved with them longer term.

The contributor group size between the non Incubated and the Chinese initiated projects are very different:

- 518 active for Chinese initiated projects
- 70 active for non-incubated projects

For email, the number of contributors for the Chinese initiated projects is high at 705 compared with 138 for the non-incubated projects.

5.3.3 Sentiment Analysis

The top mood for the Chinese initiated and the non-incubated projects groups was positivity. Trust and negativity were also present as strong sentiments.

The analysis over time for both groups shows that a positive communication style is established and is the strongest mood expressed. For the Chinese initiated group the intensity of the positivity varied a lot more than the non-Chinese projects, and at one point during February 2019, the negative mood was higher than the positive one.

The non-incubated projects also had points where negativity was higher than the positive sentiment. This includes a peak were negativity was the most dominant for a short period of time.

The overall mood rating for the non-incubated group was 63% compared with 84% for the Chinese initiated projects. Comparing the intensity of the sentiments to other projects within the ASF, the non-incubated projects are at 18% compared with 68% for the Chinese initiated projects.

In this case, the non-incubated projects level of sentiment are not as intense as other ASF projects. This shows that the Chinese initiated projects are actually more representative of the ASF mood than than the non-incubated projects.

This could be caused by the type of communication being used in the non incubated projects that lack emotion or are more focussed on specific technical tasks. It could also be that the limited number of existing contributors already understand how to work together and don't need to communicated heavily.

5.3.4 Key Phrase Analysis

Both the non-incubated and the Chinese initiated groups have similar profiles. They showed elements of standard everyday communication and technical interactions. It was interesting to see that they also showed indications of cultural expression of ASF values such as openness, collaboration and community.

There was very little cultural language being used by the non-incubated projects. They most significant thing in this comparison is that the "+1" indicator appears for the Chinese initiated projects but not for the non-incubated projects. This indicator is unique to the ASF as an indicator of consensus.

This shows that the non-incubated projects are either not using the consensus indicator, or that it used so little that it is negligible.

5.4 Incubated non-Chinese vs Non-Incubated non-Chinese

5.4.1 Pony Factors

The Pony Factor for the codebase for the non-incubated projects shows an erratic pattern with inconsistent increases and decreases potentially linked to key people being intermittently active. The incubated non-Chinese initiated projects are growing at a fast rate and still rising.

The non-incubated projects showed no difference between the Pony Factors for authorship and committership. This is probably because the people involved with the project are already experienced committers or long time mentors within the ASF.

Currently there is a large gap between the Pony Factors for authorship and committership for the incubated non-Chinese projects and the non-incubated projects.

The email Pony Factor, for the incubated non-Chinese initiated projects has grown quickly and has stabilised at 35 while the Pony Factor for the non incubated projects has decreased and is currently at 2.

5.4.2 Contributor Retention

There seems to be a good mix and flow of contributors in both the non Incubated and the incubated non-Chinese groups, with the majority of them being involved for less than a year.

With 85% of contributors with less than 2 years experience the incubated non-Chinese initiated projects show that they are successful in attracting new contributors. The non-incubated projects have more of a mix of contributors indicating that their contributors transition to being involved in the projects longer term.

The contributor group size between the non-Incubated and the incubated non-Chinese initiated projects are very different:

- 631 active for non-Chinese initiated projects
- 70 active for non-incubated projects

For email, the number of contributors for the incubated non-Chinese initiated projects is high at 1588 compared with 138 for the non-incubated projects.

5.4.3 Sentiment Analysis

The top mood for the incubated non-Chinese initiated and the non-Incubated groups was positivity. Trust and negativity were also present as strong sentiments. The analysis over time for both groups shows that the positive communication style is established and is generally the strongest mood expressed. The non incubated projects did have points where negativity was higher than the positive sentiment and includes a peak were negativity was the most dominant for a short period.

The overall mood for the non-incubated group was 63% compared with 81% for the incubated non-Chinese initiated projects. Comparing the intensity of the sentiments to other projects within the ASF, the non-incubated projects are at 14% compared with 84% for the incubated non-Chinese initiated projects.

In this case the incubated non-Chinese initiated projects are actually more representative of the ASF mood than than the non-incubated projects.

5.4.4 Key Phrase Analysis

Both the non-incubated and the incubated non-Chinese initiated groups have similar profiles. They showed elements of standard everyday communication and technical interactions. They also showed to different extents indications of cultural expression of ASF values such as openness, collaboration and community.

There was very little cultural language being used by the non-incubated projects. The most significant thing in this comparison is that the "+1" indicator appears for the incubated non-Chinese initiated projects but not for the non incubated projects. This indicator is unique to the ASF as an indicator of consensus.

This shows that the non-incubated projects are either not using the consensus indicator, or that it used so little that it is not significant.

Chapter 6

Conclusions

The purpose of this study was to gain a better understanding of the open source culture at the Apache Software Foundation and analyse a project that has a different cultural profile to see if cultural changes occur as part of the project evolution and graduation to Top Level Project.

A key part of graduating and becoming a Top Level project involves the demonstration of Apache behaviour and cultural values.

This study focuses on four Chinese contributed projects, all of which have undergone or are currently undergoing incubation at the Apache Software Foundation.

It specifically was looking to answer the following research questions:

1. What evidence can we find of cultural embedding?
2. How much of a cultural difference is there between the Chinese contributed projects and non Chinese contributed projects?
3. How successful is Apache Incubator in embedding ASF culture into the Chinese contributed projects?

It used a set of tools and indicators to create a cultural baseline based on the values and behaviours shown by the first ever ASF project, Apache HTTP Server.

A set of 8 indicators was used to create the baseline by mining the data publicly available from the ASF project archives.

For the Apache HTTP Server baseline, the following indicators were used to capture, highlight and measure:

- Pony Factors:
  - Diversity of the community, confirmation that merit is being rewarded, indication of community growth, retention of contributors
- Sentiment Analysis:
  - Dominant emotions being displayed in community interactions, communication style, overall mood of the communication (negative, positive or neutral) over time potential range of cultural elements
- Key Phrase Analysis:
  - Identifying the most common important phrases and words being used, indication of collaboration, identifies the use of unique cultural language

Twelve ASF projects were selected and divided into three groups for comparison against the Apache HTTP Server cultural baseline. The groups were also compared against each other. The groups were as follows:

- four Apache projects which have undergone incubation that have been initiated by Chinese contributors
- four Apache projects which have undergone incubation that have not been initiated by Chinese contributors
- four Apache projects which did not undergo incubation that have not been initiated by Chinese contributors

6.1 Responses to Research Questions

6.1.1 What evidence can we find of cultural embedding?

The analysis shows the following:

- the growth in Pony Factors indicate that merit is being recognised
- the contributor breakdowns show that new people are being welcomed into the project communities
- the contributor retention breakdowns show that new contributors are transitioning into longer term contributors
- cultural language is being used and this includes the "+1" consensus indicator

The key finding is that incubated projects are more proficient at using the "+1" indicator than non-incubated projects.

6.1.2 How much of a cultural difference is there between the Chinese contributed projects and non-Chinese contributed projects?

The analysis shows the following:

- the Chinese projects appeared to recognise merit at a faster level than the non Chinese projects
- the Chinese projects were more successful than the non-Chinese contributed projects in attracting new contributors to their projects
- the non-Chinese projects communicate at a higher, more frequent level than the Chinese contributed projects

- The non-Chinese projects sentiments and the intensity of the sentiment are more reflective of other ASF projects, and in comparison the Chinese contributed projects seem a little subdued
- both incubated Chinese contributed and incubated non-Chinese projects strongly adopted ASF cultural language and are frequent users of the "+1" consensus indicator

The key finding is that both the incubated Chinese and incubated non-Chinese contributed projects have strongly adopted ASF cultural behaviour and language and are frequent users of the "+1" consensus indicator.

6.1.3 How successful is Apache Incubator in embedding Apache culture in the Chinese contributed projects?

The analysis shows that Apache Incubator is very successful in embedding Apache culture into the Chinese contributed projects. The Chinese contributed projects:

- are the most proficient in recognising merit in their contributors
- are the most successful at attracting new contributors to their projects
- have adapted to use Apache cultural language
- have adopted the "+1" consensus indicator and are using it significantly

The key finding is only the incubated projects are frequent users of the "+1" consensus indicator.

6.2 Summary

Open source enables and fosters distributed development. What makes it different from general software development is that it is an environment where developers

share what they have created so that others can see what has already been done and contribute if they want to. This sharing of code is the basis of "gift culture", where the software is the gift that is being given. It is this gift of software that helps bring together and build a community. (Pan & Bonk, 2007)

This study examined Chinese contributed and non-Chinese contributed projects at the ASF and found evidence of cultural embedding. The incubated projects showed more cultural traits than the non incubated ones, in terms of mood, sentiment, contributors, and language.

The non incubated projects appear to be not as focussed on community growth, yet still manage to retain contributors. Rather than being full of experienced contributors, the non incubated projects are still attracting new people. One area of fragility is that these communities appear to be centred around a small number of people.

There are communication differences where the different project groups expressed themselves and their sentiments at different intensities. The Chinese initiated projects seemed to conform to the model of focussing on actions rather than words. So while there are apparent differences in communication style the focus on working together as a community to achieve something is the common theme.

So what of power distance and the ability to accept and expect inequalities? Actually in this case the Chinese projects are the most proficient at recognising merit which seems to imply that they understand how to use their power to encourage equality within their community of contributors.

Before each project entered incubation there was an existing community around it. Even though it may have been small, social interactions must have been occurring to enable the project to progress and survive.

On collectivism, the ASF culture and the Chinese initiated projects were probably no so far apart, as the ASF's "community over code" mantra appears to be something that was already existing within these projects. The fact that the Chinese projects are more similar to Apache HTTP Server in some areas shows that

the behaviour is occurring naturally and is already internalised.

The difference in Pony Factors for the non-Chinese initiated projects could possibly be linked to the time taken to transition behaviour from individualist thinking into collectivist or community based thinking.

Some of these project communities may have originated as corporate ones, as several corporates have created internal projects only to later open source them. The result is a transition of power as the project direction moves from corporate led to community led. So this could also be what we are seeing in some of the results in this study.

With continued growth and the ability to attract new projects and new contributors, the ASF shows that it has something unique that makes individuals want to become involved in and participate. That "something unique" is based on social interaction, culture and community.

It would be useful to be able to build on the results of this paper with further research in the following areas:

- Deeper analysis of key phrases focussing on the "+1" indicator to track this back to communities, a mailing list and a discussion thread
- Investigation to see if the initial size of a project community or Project Management Committee (PMC) affects its speed and evolution through Apache Incubator
- Investigation of the projects where corporates may be involved to analyse any differences in culture or community behaviour
- Analysis of projects with experienced mentors to see if they progress through incubation at a faster rate than those with less experienced mentors
- Analysis of non ASF related open source projects and comparison with an ASF projects to confirm if ASF values and behaviour are cultural specific / unique

- Investigation of ASF projects that have gone into decline to see if the decline could have been predicted using any of the indicators used in this paper

Appendix A

(REFERENCES)

Carillo, K., & Okoli, C. (2009). THE OPEN SOURCE MOVEMENT: A REVOLUTION IN SOFTWARE DEVELOPMENT. The Journal of Computer Information Systems, 49(2), 1-9.

Sharifian, F. (2017). Cultural Linguistics: Cultural Conceptualisations and Language. Amsterdam: John Benjamins Publishing Company.

Hofstede, G., Hofstede, G. J. & Minkov M. (2015). Cultures and organizations: Software of the mind (Rev. and expand 3rd ed.). New York: McGraw-Hill.

Zhou, Y. (2011). AGAINST INTELLECTUAL MONOPOLY: FREE SOFTWARE IN CHINA. World Review of Political Economy, 2(2), 290-306.

Pan, G., & Bonk, C. J. (2007). The Emergence of Open-Source Software in China. The International Review of Research in Open and Distributed Learning, 8(1).

Ceraso, A., & Pruchnic, J. (2011). INTRODUCTION: OPEN SOURCE CULTURE AND AESTHETICS. Criticism, 53(3), 337-375.

Caldwell-Harris, C., Kronrod, A. & Yang, J. (2013). Do more, say less: Saying "I love you" in Chinese and American cultures. Intercultural Pragmatics, 10(1), pp. 41-69. Kelty, C. M. (2004). Culture's open sources: Software, copyright, and

cultural critique. Anthropological Quarterly, 77(3), 499-506.

Incubation Policy, Apache Incubator https://incubator.apache.org/policy/incubation.html

Apache Project Maturity Model HTTP://community.apache.org/apache-way/apache-project-maturity-model.html

Incubator Proposals HTTPs://wiki.apache.org/incubator/KylinProposal HTTPs://wiki.apach HTTPs://wiki.apache.org/incubator/SkyWalkingProposal HTTPs://wiki.apache.org/incubator/

Allen, D. (2003). Cultural conundrum. Nursing Standard (through 2013), 17(43), 16-8.

Seidler, Katie. Crime, Culture and Violence : Understanding How Masculinity and Identity Shapes Offending, Australian Academic Press, 2010. ProQuest Ebook Central,

Rose, M. (2007). THE TRANSMISSION OF CULTURE IN A GLOBALIZED WORLD: THE CHALLENGE OF NEW TECHNOLOGIES. Canadian Issues, , 55-57.

Murray, P. (2006). Open sources 2.0: The continuing evolution. Weybridge: Birchley Hall Press.