Ecosystems in GitHub and a Method for Ecosystem Identification using Reference Coupling

Kelly Blincoe, Francis Harrison and Daniela Damian
Software Engineering Global InterAction Lab
University of Victoria
Victoria, BC, Canada
Email: kblincoe@acm.org, francish@uvic.ca, danielad@uvic.ca

***Abstract*—Software projects are not developed in isolation. Recent research has shifted to studying software ecosystems, communities of projects that depend on each other and are developed together. However, identifying technical dependencies at the ecosystem level can be challenging. In this paper, we propose a new method, known as reference coupling, for detecting technical dependencies between projects. The method establishes dependencies through user-specified cross-references between projects. We use our method to identify ecosystems in GitHub-hosted projects, and we identify several characteristics of the identified ecosystems. We find that most ecosystems are centered around one project and are interconnected with other ecosystems. The predominant type of ecosystems are those that develop tools to support software development. We also found that the project owners' social behaviour aligns well with the technical dependencies within the ecosystem, but project contributors' social behaviour does not align with these dependencies. We conclude with a discussion on future research that is enabled by our reference coupling method.**

I. Introduction

Software ecosystems, defined as "a collection of software projects which are developed and which co-evolve together in the same environment" [1], have become an area of interest in recent research. Since software projects are not typically developed in isolation, studying a software project without examining its surrounding ecosystem is incomplete. Thus, analysis of software ecosystems has emerged as a novel new research area in recent years.

Projects in an ecosystem depend on one another [1]. The technical dependencies that exist between projects define the structure of the ecosystem [2]. Thus, identifying technical dependencies between software projects is a useful way to identify ecosystems. However, identifying technical dependencies between projects on a large scale has proven to be difficult [3]. Existing static dependency analysis approaches do not identify dependencies across projects. Methods for extracting dependencies from a project's source code have been proposed [2], [3], [4], but they require large amounts of memory and computation time [5]. Thus, they cannot be employed across a large set of projects. Other methods [1], [6], [7], [8], [9] avoid analyzing source code by obtaining technical dependency information from a project's configuration files, but this information is not always available or accurate. Without a way to establish dependencies between projects, researchers cannot fully understand a project's ecosystem.

To identify technical dependencies, we turn to cross-references on GitHub, a social code hosting service. GitHub encourages collaboration between users both within and across projects through its transparent interface and built-in social features. With GitHub Flavored Markdown[1], when a user cross-references another repository in a pull request, issue or commit comment a link to the other repository is automatically created. This introduces a way for developers to bring awareness across repositories. In this paper, we investigate whether these cross-references indicate a technical dependency between the two repositories by examining a set of these cross-references. We found that the cross-references are a good conceptualization of technical dependencies between projects, and we highlight several common types of technical dependencies seen in these cross-reference comments. We call our method for identifying technical dependencies reference coupling.

With an ability to identify technical dependencies between a large number of projects, ecosystems of densely connected projects can be identified. We use a popular community detection algorithm [10] to identify ecosystems across all GitHub-hosted projects. Ecosystem identification is important to help developers understand how their tasks fit into the big picture and who they need to coordinate their changes with at the ecosystem level [1], [11]. For open source ecosystems, it is also important for attracting new contributors [1] since project's within an ecosystem are more likely to attract attention. In this paper, we analyze the ecosystems identified using our methods on GitHub-hosted projects and identify a set of properties that characterize the ecosystems. Conway Law. To complement our focus on technical dependencies within ecosystems, we also investigate the social behaviour of project owners/contributors in relation to these ecosystems. Our analysis was guided by the following research questions:

RQ1: Do cross-references to other projects in issue, pull request, and commit comments indicate the existence of a technical dependency between the two projects?

RQ2: What ecosystems exist across GitHub-hosted projects and what is their structure?

RQ3: Do the project owners' and contributors' social behaviours align with the technical dependencies?

[1]https://help.github.com/articles/github-flavored-markdown/

978-0-7695-5594-2/15 $31.00 © 2015 IEEE
DOI 10.1109/MSR.2015.26

The rest of the paper is structured as follows: Our research methods and results are presented in Section II for RQ1, Section III for RQ2, and Section IV for RQ3. In Section V, we summarize our findings and discuss open questions for future research. Section VI provides an overview of related work in software ecosystems and dependency conceptualizations. We provide a brief conclusion in Section VII.

II. Reference Coupling

RQ1: Do cross-references to other projects in issue, pull request, and commit comments indicate the existence of a technical dependency between the two projects?

A. Research Method

We studied the cross-references made to other repositories in comments on issues, pull requests and commits. We obtained data from the GHTorrent [12] project, which provides a mirror of the GitHub API data. GHTorrent obtains its data by monitoring and recording GitHub events as they occur. We used the MySQL 2014-04-02 dataset to obtain information on the projects. This dataset contains data on 2,399,526 repositories, 3,426,046 users, and their events - including commits, issues, pull requests and comments. Since the MySQL database contains only the first 256 characters of comments, we obtained all comments from GHTorrent's main MongoDB server in May 2014. The MongoDB contains the full text of all comments.

Cross-references follow the pattern User/Project#Num (e.g. rails/rails#123) or User/Project@SHA. We performed pattern matching on all comments in the GHTorrent database to identify these cross-references. We define a project as a repository and all of its forks as recommended by [13]. Since we are interested only in relationships between projects, we filtered the cross-references to ignore those where one project is a fork of the other project. We identified 89,784 comments with a cross-reference to another project.

To answer RQ1 and verify that these cross-references are a valid conceptualization of dependencies, we examined 198 random comments which cross-referenced another project. We classified a comment as a technical dependency if the comment described a work dependency, either direct or indirect, between the two projects.

We examined the comments classified as technical dependencies to identify the types of dependencies that exist through these cross-reference relationships. We used a grounded theory approach to identify types of dependencies [14]. We conducted open coding on the cross references, grouping conceptually similar comments into categories. We stopped when we achieved saturation after 49 comments. One person, familiar with software development practices, performed the qualitative coding.

B. Results

Most of the cross-reference comments (90%) were classified as technical dependencies. The remaining 10% of comments did not contain enough details for proper classification. Many of these comments simply referenced another repository using the pattern as described above with no additional text provided.

We identified two main types of dependencies:

Dependency between the two projects. The most common type of dependency found was a direct technical dependency between the two projects. An example of a direct technical dependency is when an issue created in one project depends on a fix/update in another project. Another example is when a project needs to be updated based on changes made in another project.

Below we provide three examples of cross-reference comments that are indicative of direct technical dependencies between the two projects. Project names follow the pattern user/repository where user is the owner's GitHub login and repository is the name of the project repository.

Issue #449 on the sensu/sensu project describes an issue that is the result of the interaction between the sensu/sensu code and the ruby-amqp/amq-client library. The comment references a commit on the ruby-amqp/amq-client that fixes the issue.

> *"I verified that the problem is still the one referenced in ruby-amqp/amq-client#14. This fix is not merged with amq-client's '0.9.x-stable' branch. This is why I am still hitting it. The commit ruby-amqp/amq-client@60f1c59 is the fix but it resides only in the master branch."*

Issue #8 on the tsujigiri/axiom project notes that changes must be made to the code base to allow an upgrade to the latest release of the ninenines/cowboy project.

> *"Upgrade Cowboy: After Cowboy 0.6.1 Cowboy's http_req record was made opaque and can not be used directly anymore. I didn't really have the time yet to look into it, but it looks like we just need to remove all references to the record from the documentation and add directions on how to access cowboy_req:req() via the cowboy_req functions. See ninenines/cowboy#266 and ninenines/cowboy#267."*

Commit 81bbbec21c04b6392f6892f7735243387d295337 on the joyent/node project closes isaacs/node-graceful-fs issue #6, which describes a problem in the isaacs/node-graceful-fs code stemming from the use of joyent/node. GitHub allows automatic closure of issues through commit comments, even when the commit is in a different repository[2].

> *"This fixes isaacs/node-graceful-fs#6."*

Both projects depend on a third project. We also identified some cases where the comments describe a dependency on a third project that is not cross-referenced. For example, everzet/capifony's pull request #376 cross-references composer/composer's issue #1453, but the problem stems from the use of the symfony/symfony project. After identifying the source of the problem, a new issue (#411) is created on the

[2]https://github.com/blog/1439-closing-issues-across-repositories

everzet/capifony project that identifies the changes that need to be made to the way the symfony environment is set so that the composer/composer code executes correctly.

> *"As described in #376 capifony should execute composer with the right symfony environment set. Currently, with `--no-scripts` option removed in #376, composer is always executing symfony scripts with default dev environment."*

Cross-references to other repositories appearing in GitHub comments, therefore, do indicate the existence of a technical dependency. We call this conceptualization of dependencies between projects reference coupling.

III. GitHub Ecosystems

RQ2: What ecosystems exist across GitHub-hosted projects and what is their structure?

A. Research Method

We constructed a network of the technical dependency relationships established through reference coupling as described in Section II. The *Dependency Network* is defined as a directed graph $G_d =< V, E >$. The set of vertices, denoted by V, is all GitHub projects involved in at least one cross-reference. There are 18,533 projects in this set. The set of edges, denoted by E, is a set of node pairs $E(V) = \{(x, y)|x, y \in V\}$. If the project represented by node x_i cross-referenced the project represented by node y_j, there is a directed edge from x_i to y_j. The weight of each edge is the count of cross-references for the pair of projects. We filtered the edges to only consider dependencies between nodes if the pair of projects have been cross-referenced two or more times to capture only the stronger dependencies.

To identify ecosystems across projects hosted on GitHub, we used the popular Louvain community detection method [10] on the Dependency Network. The Louvain method is a greedy optimization method that aims to partition a network into communities of densely connected nodes and optimize the modularity of the network. Modularity is defined as "the number of edges falling within [communities] minus the expected number in an equivalent network with edges placed at random [15]." The Louvain method is comprised of two steps. It first optimizes modularity locally by looking for small communities. Then it aggregates the nodes in each small community and builds a new network with these aggregated nodes. It iterates on these two steps until the modularity is maximized. The Louvain method outperforms all other community detection methods in terms of both the modularity that is achieved and the computation time [10].

High modularity scores indicate that there are dense connections within the communities but sparse connections across communities, showing that an optimal solution has been found. When high modularity scores are obtained, the communities have significant real-world meaning [10]. In our network, the identified communities represent sets of projects densely connected by technical dependencies. Since dependencies that exist between projects define the structure of an ecosystem [2], these communities represent software ecosystems.

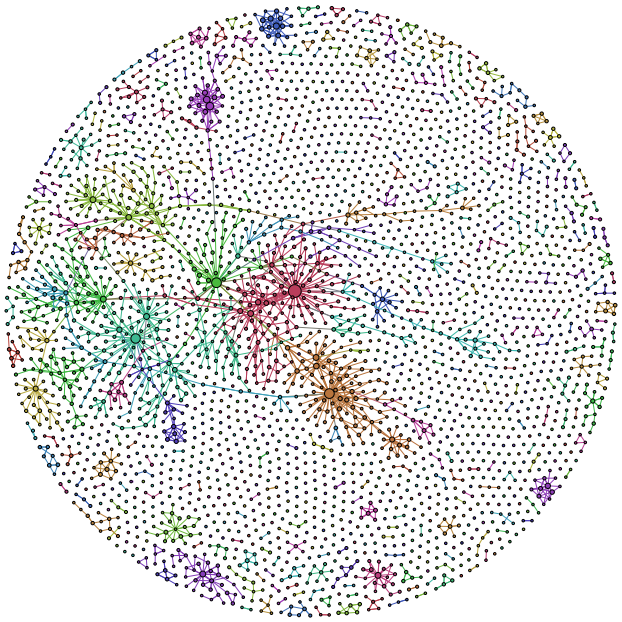

Fig. 1. All GitHub projects with cross-references. The largest connected component (or giant component) is easily identified as the well-connected subgraph appearing in the center of the graph.

To identify properties of the identified ecosystems, we analyzed visualizations of the Dependency Network. We used the Gephi [16] graphing tool to create visualizations. We inspected visualizations of the network to visually identify patterns. We triangulated the patterns visible in the network with statistics of the network. We also examined the type of projects prominent in each ecosystem.

B. Results

Of the 18,533 projects in the Dependency Network, 10,484 (57%) are a part of the largest connected component (commonly referred to as the giant component [17]), which is the largest subgraph in which every node is connected to every other node by some path. The connected components isolated from the giant component are primarily comprised of *same owner* communities in which all nodes in the connected component are projects owned by the same GitHub user or organization. For example, the second largest connected component is comprised of 65 nodes, of which, all but two are owned by GitHub user deathcap[3]. Figure 1 shows the full Dependency Network, though for visibility we only display nodes with degree of 3 or greater. As visible on the graph, most of the nodes isolated from the giant component are connected to only a small number of nodes. In fact, 75% of nodes not in the giant component are connected to only one other node.

Since we are most interested in studying the popular GitHub ecosystems, we focus our analysis on the interconnected part

[3]https://github.com/deathcap

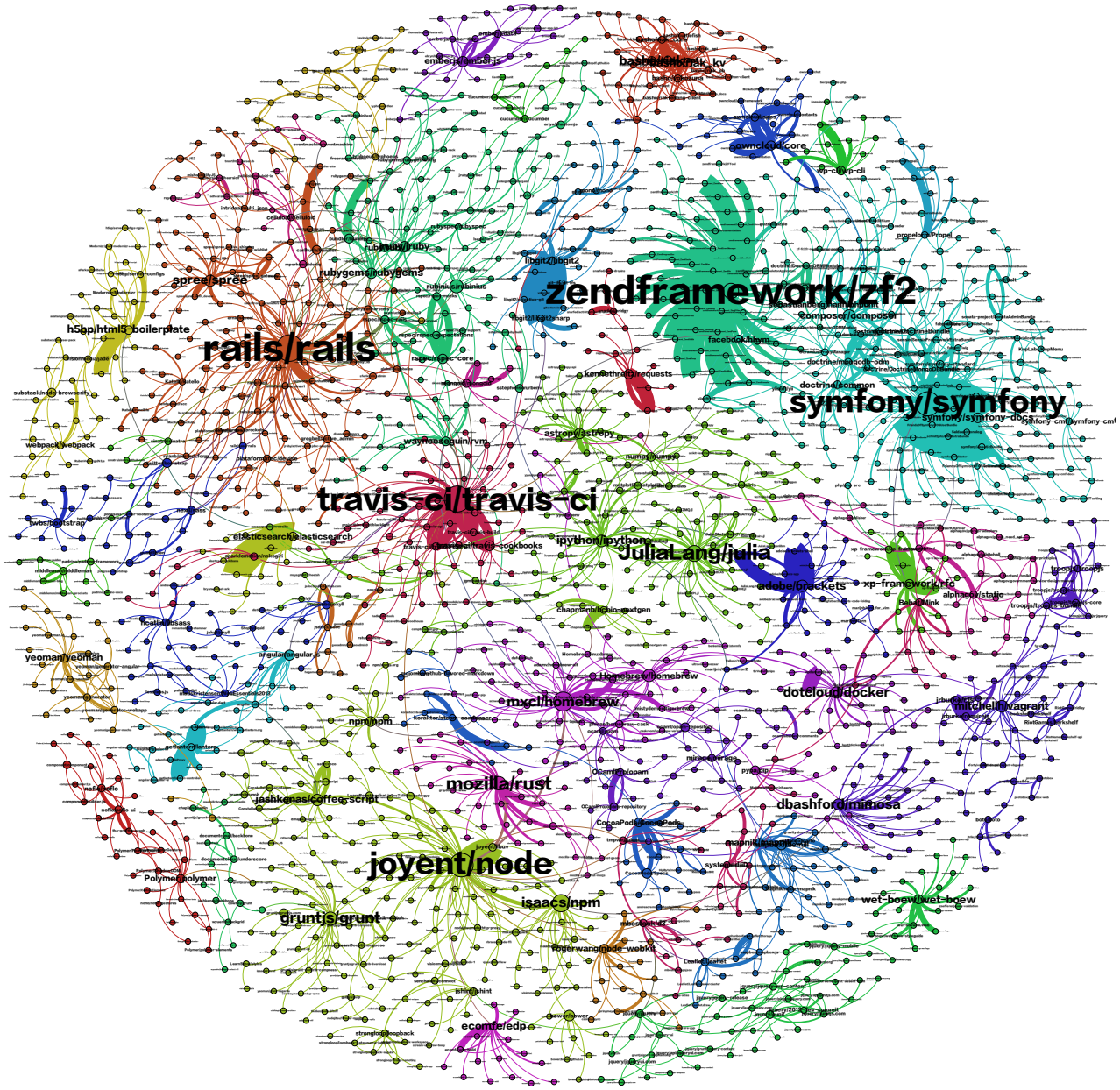

Fig. 2. Ecosystems in the largest connected component of GitHub-hosted projects. Project names follow the pattern user/repository where user is the owner's GitHub login and repository is the name of the project repository.

of the network or the giant component. Figure 2 shows the giant component. The color of the nodes represent communities as detected by the Louvain method. We obtained a modularity score of 0.913 (out of a possible range of 0 to 1). This high modularity score indicates that the detected communities are much more tightly connected by technical dependencies than would appear in a random graph.

There were 43 ecosystems identified in this network. Nodes are sized according to their authority to display the nodes that are more prominent in each ecosystem. When a node has a high number of cross-reference relationships pointing to it, it has a high authority value [18]. Table I shows the most well-connected project node (highest Authority value) in each of the ecosystems.

C. Properties of GitHub Ecosystems

Ecosystems revolve around one central project. As depicted in Figure 2, each ecosystem appears to revolve around one main project. In Table I, the most well-connected project node in each ecosystem is listed along with a description

TABLE I
ECOSYSTEMS IN GITHUB. DETAILS OF THE MOST WELL-CONNECTED NODE IN EACH ECOSYSTEM.

Project	Description	Stars	Ecosystem Size	Degree (in,out)
joyent/node	Framework	39,373	10.08%	69 (53,16)
symfony/symfony	Framework	10,985	8.46%	93 (53,40)
rails/rails	Framework	29,744	7.92%	93 (65,28)
JuliaLang/julia	Programming Language	5,531	6.74%	51 (35,16)
rubygems/rubygems	Package Manager	1,304	6.04%	22 (14,8)
mxcl/homebrew	Package Manager	13,723	3.94%	48 (21,27)
zendframework/zf2	Framework	5,841	3.88%	72 (65,7)
travis-ci/travis-ci	Development Tool (Continuous Integration Platform)	3,693	3.50%	70 (54,16)
wet-boew/wet-boew	Framework	688	3.34%	19 (15,4)
twbs/bootstrap	Framework	41,828	3.29%	9 (9,0)
dbashford/mimosa	Development Tool (Browser development)	472	2.43%	25 (20,5)
h5bp/html5-boilerplate	Framework	31,926	2.37%	19 (15,4)
mitchellh/vagrant	Framework	9,274	2.10%	23 (15,8)
libgit2/libgit2	Library	5,161	2.05%	20 (11,9)
Behat/Mink	Development Tool (Testing)	673	1.99%	13 (9,4)
OCamlPro/opam	Package Manager	118	1.89%	9 (8,1)
basho/riak	Database	2,520	1.83%	27 (18,9)
Polymer/polymer	Library	8,787	1.83%	16 (11,5)
mapnik/mapnik	Development Tool (Toolkit for developing mapping applications)	1,003	1.78%	20 (12,8)
mozilla/rust	Programming language	5,604	1.78%	36 (29,7)
alphagov/static	Other (GOV.UK static files/resources)	67	1.73%	13 (10,3)
adobe/brackets	Development Tool (code editor)	23,921	1.46%	26 (16,10)
CocoaPods/CocoaPods	Development Tool (dependency manager)	5,711	1.46%	14 (9,5)
yeoman/yeoman	Development Tool (web development tools)	7,246	1.46%	18 (13,5)
angular/angular.js	Framework	42,950	1.40%	12 (8,4)
dotcloud/docker	Development Tool (application container engine)	14,270	1.35%	24 (19,5)
emberjs/ember.js	Framework	14,185	1.29%	20 (12,8)
owncloud/core	Other (personal cloud storage tool)	3,222	1.19%	26 (13,13)
typhoeus/typhoeus	Library	2,465	1.19%	6 (4,2)
facebook/hhvm	Other (Virtual machine)	11,506	1.08%	15 (10,5)
celluloid/celluloid	Framework	2,855	0.86%	9 (6,3)
xp-framework/rfc	Framework	0	0.86%	16 (14,2)
rogerwang/node-webkit	Framework	19,737	0.86%	16 (11,5)
ecomfe/edp	Development Tool (front-end development platform)	264	0.86%	18 (15,3)
kennethreitz/requests	Library	13,812	0.81%	13 (10,3)
documentcloud/underscore	Library	7,135	0.81%	6 (4,2)
middleman/middleman	Development Tool (website generator)	4,179	0.75%	8 (5,3)
elasticsearch/elasticsearch	Other (search and analytics tool)	10,700	0.70%	11 (11,0)
chapmanb/bcbio-nextgen	Other (RNA-seq analysis tool)	173	0.59%	10 (9,1)
wp-cli/wp-cli	Development Tool (command line interface for WordPress)	1,968	0.59%	13 (9,4)
cucumber/cucumber	Development Tool (Testing)	5,142	0.49%	7 (4,3)
jsdoc3/jsdoc	Development Tool (API documentation generator)	2,909	0.49%	6 (3,3)
propelorm/Propel	Development Tool (Object-Relational Mapping)	893	0.49%	7 (7,0)

of the project, the number of stars the project has, the size of the associated ecosystem, and the node's degree. Each of these projects has a higher in-degree than out-degree with the exception of the mxcl/homebrew project. On the other hand, low-degree project nodes are four times as likely to be dependent on another project than they are to have a project depend on them. This shows that ecosystems are being formed around a central project with the other projects in the ecosystem mostly depending on that central project. This results in a star pattern. The twbs/bootstrap ego network (Figure 3) clearly depicts this pattern within the graph.

Predominant type of ecosystems is software development support. Interestingly, nearly all of the ecosystems are centered around projects whose purpose is to support software development, such as frameworks, libraries and programming languages. In fact, of the 43 ecosystems, there are only 5 whose purpose is not to support software development. There are 13 frameworks, 5 libraries, 3 package managers, 2 programming languages, 1 database, and 14 other tools that support software development like a testing tool, a continuous integration platform, and an API documentation generator.

Ecosystems are interconnected. The graph in Figure 1 shows two types of communities that occur in GitHub-hosted projects, those that are part of the largest connected component and those that are isolated from the largest connected component. The majority of project nodes, 10,484 or 57%, are involved in the largest connected component, indicating that many ecosystems are connected to each other across the projects in our Dependency Network. The next biggest connected component in the graph is only 65 nodes indicating that the ecosystems that are isolated are small and have not attracted public attention.

Figure 2 displays the interconnected part of the network, and the connections between the ecosystems are apparent. As an

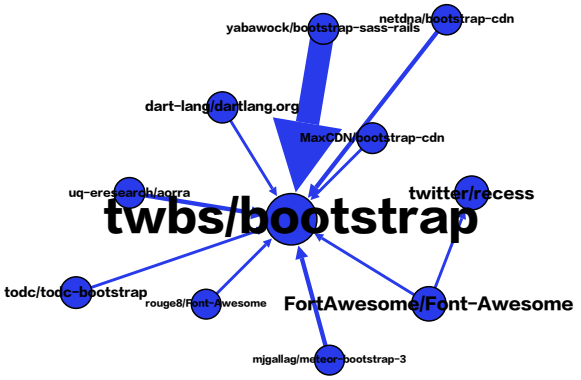

Fig. 3. twbs/bootstrap Ego Network. Portraying a sample star pattern in the network.

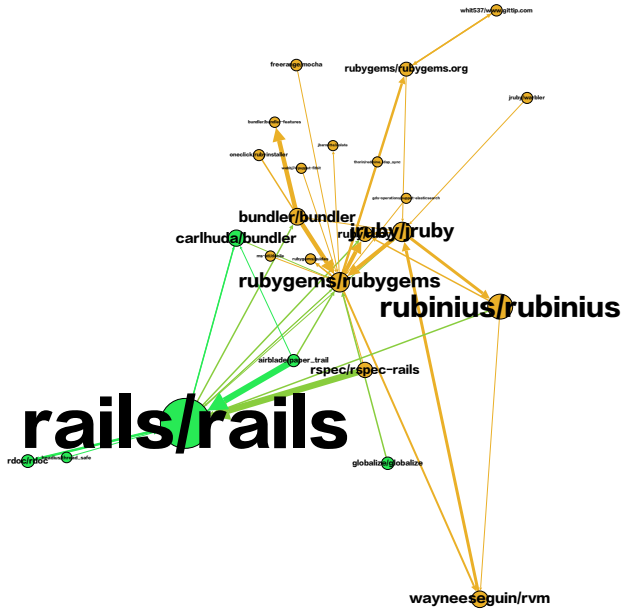

Fig. 4. rubygems/rubygems Ego Network. Portraying connections between ecosystems.

example, Figure 4 shows the rubygems/rubygems ego network, clearly depicting its connection to the rails/rails project. This is not surprising, since the rubygems project is a package management framework for the Ruby programming language and rails/rails is a web application framework written in Ruby. There is a direct connection between the rubygems/rubygems and rails/rails nodes. In addition, there are projects, like carlhuda/bundler and airblade/paper_trail, which connect the two projects.

IV. Social Behaviour of Project Owners/Contributors

RQ3: Do the project owners' and contributors' social behaviours align with the technical dependencies?

A. Research Method

To complement our investigation of technical dependencies and connectedness of projects in Github, we also sought to understand the social behaviour of project owners/contributors in relation to these ecosystems. We studied two of GitHub's social relationships, following users and starring projects. On GitHub, users can *follow* other users to receive notification on their activity and *star* a repository to bookmark it or indicate interest in the project. To understand how the social behaviour of project owners/contributors relates to the identified ecosystems, we examine the alignment between social and technical connections between the projects.

To answer our research question, we construct project-to-project networks based on the following and starring activity of the project owners and contributors. We ran correlations between the edge weights in the Dependency Network with the edge weights in each social network to determine if there is a relationship between the technical dependencies and the social connections. Pearson correlations were used since the data was normally distributed.

Project Owners. We constructed two networks using the following and starring relationships by considering the actions of the project owners. The *Owner Stars Network*, $G_{os} =< V, E >$, and the *Owner Follows Network*, $G_{of} =< V, E >$, are both undirected graphs whose set of vertices is all GitHub projects involved in at least one cross-reference. For the Owner Follows Network, there is an edge from nodes x_i to y_j if the owner of project x_i follows the owner of project y_j. There is an edge from x_i to y_j in the Owner Stars Network if an owner of any project in our dataset has starred both project x_i and project y_j.

To compare the social connections with the technical dependencies, we compare the edge weights of these two networks with the edge weights of the Dependency Network described earlier. The edge weights of these three networks represent the following:

- Dependency Network G_d: Number of technical dependencies, measured through reference coupling, between the two project nodes.
- Owner Follows Network G_{of}: 0 if neither project owner follows the other, 1 if one project owner follows the other project owner, and 2 if both project owners follow each other.
- Owner Stars Network G_{os}: Number of project owners who have starred both projects.

Project Contributors. We constructed two additional networks using these following and starring relationships by considering the actions of the project contributors (users who have made commits on the project or are members of the project). The *Contributor Stars Network*, $G_{cs} =< V, E >$, and the *Contributor Follows Network*, $G_{cf} =< V, E >$, are also undirected graphs whose set of vertices is all GitHub projects involved in at least one cross-reference. The Contributor Follows Network has an edge from nodes x_i to y_j if a contributor