End-User Programmers and their Communities: An Artifact-based Analysis

Kathryn T. Stolee, Sebastian Elbaum, and Anita Sarma
University of Nebraska–Lincoln

{kstolee, elbaum, asarma}@cse.unl.edu

September 22, 2011

This work is supported by the NSF GRFP under CFDA#47.076, NSF Award #0915526, and AFOSR Award #9550-10-1-0406.
End User Programmers

People who engage in programming activities to support their hobbies and work.
End User Programmers

People who engage in programming activities to support their hobbies and work.

<table>
<thead>
<tr>
<th>Professionals</th>
<th>End Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number in U.S.</td>
<td>3 million</td>
</tr>
<tr>
<td>Typical Education</td>
<td>C.S. Degree</td>
</tr>
<tr>
<td>Role of Programming</td>
<td>It’s their job</td>
</tr>
<tr>
<td></td>
<td>13 million</td>
</tr>
<tr>
<td></td>
<td>Other Degree</td>
</tr>
<tr>
<td></td>
<td>It supports their job</td>
</tr>
</tbody>
</table>
Many Domains and Applications

Web Mashups:

Educational Games:

Scientific Computing:

Yet we know little about how the repositories are utilized.
Many Domains and Applications

Web Mashups: Educational Games: Scientific Computing:

<table>
<thead>
<tr>
<th>Environment</th>
<th>Yahoo! Pipes</th>
<th>Scratch</th>
<th>MATLAB</th>
</tr>
</thead>
<tbody>
<tr>
<td># Artifacts</td>
<td>100,000</td>
<td>700,000</td>
<td>13,717</td>
</tr>
<tr>
<td># Participants</td>
<td>90,000</td>
<td>500,000</td>
<td>5,356</td>
</tr>
</tbody>
</table>
Many Domains and Applications

Web Mashups: 
Educational Games: 
Scientific Computing:

<table>
<thead>
<tr>
<th>Environment</th>
<th>Yahoo! Pipes</th>
<th>Scratch</th>
<th>MATLAB</th>
</tr>
</thead>
<tbody>
<tr>
<td># Artifacts</td>
<td>100,000</td>
<td>700,000</td>
<td>13,717</td>
</tr>
<tr>
<td># Participants</td>
<td>90,000</td>
<td>500,000</td>
<td>5,356</td>
</tr>
</tbody>
</table>

... yet we know little about how the repositories are utilized
Empirical Study Details

- Research Goal
- Study Context
- Research Questions
- Variables and Metrics
- Methods
- Results
Research Goal

To better understand *end-user programmer* communities

- Learn how communities and artifact repositories evolve
- Uncover needs for support in: development, maintenance, search, program understanding, ...
Empirical Study Details

**Goal:** To better understand *end-user programmer* communities

- Research Goal
- **Study Context**
- Research Questions
- Variables and Metrics
- Methods
- Results
Why Mashup Communities?

Web Mashups

Applications that compose and manipulate existing data sources or services to create new data or service.

Why study mashups?

- Many environments (e.g., Apatar, DERI Pipes, IBM Mashup Center, Kivati, Yahoo! Pipes, . . .)
- Potential impact (many users, growth)
Why Mashup Communities?

Web Mashups

Applications that compose and manipulate existing data sources or services to create new data or service.

Why study mashups?

- Many environments (e.g., Apatar, DERI Pipes, IBM Mashup Center, Kivati, Yahoo! Pipes, ...)  
- Potential impact (many users, growth)
End-User Programmers and their Communities

Web Mashups

About Yahoo! Pipes

This example mashup fetches and filters news from news.google.com.

Information page shows the pipe output and descriptive information.
About Yahoo! Pipes

Clicking **Publish** adds the pipe to the public repository.
End-User Programmers and their Communities

Empirical Study

Web Mashups

About Yahoo! Pipes

Clicking Edit Source loads the Pipes Editor
About Yahoo! Pipes

- Visual mashup creation environment
- Within a browser
- Drag and drop interface
About Yahoo! Pipes

- Visual mashup creation environment
- Within a browser
- Drag and drop interface
About Yahoo! Pipes

- Visual mashup creation environment
- Within a browser
- Drag and drop interface
About Yahoo! Pipes

- Visual mashup creation environment
- Within a browser
- Drag and drop interface
Empirical Study Details

**Goal:** To better understand *end-user programmer* communities

- Research Goal
- Study Context
- **Research Questions**
- Variables and Metrics
- Methods
- Results
Research Questions

**RQ1:** What are the characteristics of Yahoo! Pipes community?

1a,b: author attrition and author contributions

1c: artifact sharing, abstraction, complexity, and degree of overlap among pipes in the repository
Research Questions

**RQ1:** What are the characteristics of Yahoo! Pipes community?
   1a,b: author attrition and author contributions
   1c: artifact sharing, abstraction, complexity, and degree of overlap among pipes in the repository

**RQ2:** How do pipe attributes change as authors gain experience?
   2a: experience measured by time
   2b: experience measured by total contributions
Research Questions

**RQ1:** What are the characteristics of Yahoo! Pipes community?
   - 1a,b: author attrition and author contributions
   - 1c: artifact sharing, abstraction, complexity, and degree of overlap among pipes in the repository

**RQ2:** How do pipe attributes change as authors gain experience?
   - 2a: experience measured by time
   - 2b: experience measured by total contributions

**RQ3:** What are the characteristics of most prolific authors?
   - 3a: author activities
   - 3b: author skills
   - 3c: awareness of the community
Research Questions

**RQ1:** What are the characteristics of Yahoo! Pipes community?
   1a,b: author attrition and author contributions
   1c: artifact sharing, abstraction, complexity, and degree of overlap among pipes in the repository

**RQ2:** How do pipe attributes change as authors gain experience?
   2a: experience measured by time
   2b: experience measured by total contributions

**RQ3:** What are the characteristics of most prolific authors?
   3a: author activities
   3b: author skills
   3c: awareness of the community
Empirical Study Details

**Goal:** To better understand *end-user programmer* communities

- Research Goal
- Study Context
- Research Questions
- **Variables and Metrics**
- Methods
- Results
### Study Details

<table>
<thead>
<tr>
<th>Concept to Capture</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>artifact sharing/impact</td>
<td>popularity</td>
</tr>
<tr>
<td>abstraction</td>
<td>configurability</td>
</tr>
<tr>
<td>complexity</td>
<td>size</td>
</tr>
<tr>
<td>overlap of artifacts in repository</td>
<td>diversity</td>
</tr>
</tbody>
</table>
Study Details

Variables: size, configurability, popularity, diversity

Pipe Source

Pipe Information

Search EZTV for TV Shows
Searches EZTV’s RSS for TV Shows that you specify and returns a RSS feed. Season and Episode Number not required.
Pipe Web Address: http://pipes.yahoo.com/pipes/pipe.info?_id=57d3c38b35160aeed07cfaba6d01b0229

Pipe Output
**Variables:** size, configurability, popularity, diversity

**Pipe Source**

1. Name/Show
2. Season
3. EpisodeNumber
4. URL Resolver
5. Fetch Feed
6. Pipe Output

**6 modules**

**Significance:** Size is related to complexity
Study Details

Variables: size, configurability, popularity, diversity

Pipe Source

Pipe Information

Search EZTV for TV Shows
Searches EzTv's RSS for TV Shows that you specify and returns a RSS feed. Season and Episode Number not required.
Pipe Web Address: http://pipes.yahoo.com/pipes/pipemetadata?_id=57d3c38b35163aee07cafaa8d1b0229

3 modules

Significance: Configurability is related to abstraction and language mastery
End-User Programmers and their Communities

Empirical Study

Metrics

Study Details

Variables: size, configurability, popularity, diversity

Pipe Source

Pipe Information

Search EZTV for TV Shows
Searches EZTV's RSS for TV Shows that you specify and returns a RSS feed. Season and Episode Number not required.
Pipe Web Address: http://pipes.yahoo.com/pipes/pipe.info?_id=57d3c38b35160aee07cfaba6d01b0229

3 modules

Significance: Configurability is related to abstraction and language mastery
Study Details

**Variables:** size, **configurability**, popularity, diversity

**Pipe Source**

**Pipe Information**

Search EZTV for TV Shows

Searches EZTV's RSS for TV Shows that you specify and returns a RSS feed. Season and Episode Number not required.

Pipe Web Address: http://pipes.yahoo.com/pipes/pipe.info?_id=56d3abc3cfb1b33ee07cfba8b0229

Significance: Configurability is related to abstraction and language mastery
Variables: size, configurability, popularity, diversity

Significance: Configurability is related to abstraction and language mastery
Study Details

**Variables:** size, configurability, **popularity**, diversity

**Pipe Source**

![Pipe Source Diagram]

**Pipe Information**

**Search EZTV for TV Shows**

Searches EzTv's RSS for TV Shows that you specify and returns a RSS feed. Season and Episode Number not required.

Pipe Web Address: [http://www.ezrasn.com/pipes/pipesinfo?_id=57d3c3b3516a3ee07c0ba8d10b0229](http://www.ezrasn.com/pipes/pipesinfo?_id=57d3c3b3516a3ee07c0ba8d10b0229)

![Pipe Information Diagram]

190 clones

**Significance:** Popularity is related to impact on community
End-User Programmers and their Communities

Empirical Study

Study Details

**Variables:** size, configurability, popularity, **diversity**

Pipe Source

1. Same structure, fields, content
2. Same structure, field counts
3. Same structure
4. Same bag of modules
5. Same set of modules
6. Same type bag
7. Same size
8. No match

**Significance:** Diversity is related to contribution novelty
Study Details

**Variables:** size, configurability, popularity, *diversity*

**Pipe Source**

1. Same structure, fields, content
2. Same structure, field counts
3. **Same structure**
4. Same bag of modules
5. Same set of modules
6. Same type bag
7. Same size
8. No match

**Significance:** Diversity is related to contribution novelty
Study Details

Variables: size, configurability, popularity, diversity

Significance: Diversity is related to contribution novelty
Study Details

Variables: size, configurability, popularity, **diversity**

Pipe Source

1. Same structure, fields, content
2. Same structure, field counts
3. Same structure
4. Same bag of modules
5. Same set of modules
6. Same type bag
7. Same size
8. No match

Significance: Diversity is related to contribution novelty
**Study Details**

**Variables:** size, configurability, popularity, **diversity**

**Pipe Source**

1. Same structure, fields, content
2. Same structure, field counts
3. Same structure
4. Same bag of modules
5. Same set of modules
6. Same type bag
7. Same size
8. No match

**Significance:** Diversity is related to contribution novelty
Study Details

**Variables:** size, configurability, popularity, **diversity**

**Pipe Source**

**Significance:** Diversity is related to contribution novelty
Study Details

Variables: size, configurability, popularity, diversity

Pipe Source

5 Same set of modules

Significance: Diversity is related to contribution novelty
Study Details

**Variables:** size, configurability, popularity, diversity

**Pipe Source**

1. Same structure, fields, content
2. Same structure, field counts
3. Same structure
4. Same bag of modules
5. Same set of modules
6. Same type bag
7. Same size
8. No match

**Significance:** Diversity is related to contribution novelty
Empirical Study Details

**Goal:** To better understand *end-user programmer* communities

- Research Goal
- Study Context
- Research Questions
- Variables and Metrics
- **Methods**
- Results
Data Collection
Data Collection

Artifacts: 32,887
Data Collection

Artifacts: 32,887
Authors: 20,313
End-User Programmers and their Communities

Empirical Study

Study Methods

Data Collection

Artifacts: 32,887
Authors: 20,313

Threats: public repository offers limited visibility (internal); sampling bias (external); generalizability to other domains (external)
Empirical Study Details

**Goal:** To better understand *end-user programmer* communities

- Research Goal
- Study Context
- Research Questions
- Variables and Metrics
- Methods
- **Results**
Research Questions

**RQ1:** What are the characteristics of Yahoo! Pipes community?

1a,b: author attrition and author contributions

1c: artifact sharing, abstraction, complexity, and degree of overlap among pipes in the repository
RQ1: Characteristics of Yahoo! Pipes Community

**Summary**

<table>
<thead>
<tr>
<th>Metric</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>8.20 modules per pipe</td>
</tr>
<tr>
<td>Configurability</td>
<td>0.65 modules per pipe</td>
</tr>
<tr>
<td>Popularity</td>
<td>5.67 clones per pipe</td>
</tr>
<tr>
<td>Diversity</td>
<td>3.62 cluster level</td>
</tr>
</tbody>
</table>
RQ1: Characteristics of Yahoo! Pipes Community

<table>
<thead>
<tr>
<th>Metric</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>8.20 modules per pipe</td>
</tr>
<tr>
<td>Configurability</td>
<td>0.65 modules per pipe</td>
</tr>
<tr>
<td>Popularity</td>
<td>5.67 clones per pipe</td>
</tr>
<tr>
<td>Diversity</td>
<td>3.62 cluster level</td>
</tr>
</tbody>
</table>

34% of pipes are configurable
RQ1: Characteristics of Yahoo! Pipes Community

<table>
<thead>
<tr>
<th>Metric</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>8.20 modules per pipe</td>
</tr>
<tr>
<td>Configurability</td>
<td>0.65 modules per pipe</td>
</tr>
<tr>
<td><strong>Popularity</strong></td>
<td><strong>5.67 clones per pipe</strong></td>
</tr>
<tr>
<td>Diversity</td>
<td>3.62 cluster level</td>
</tr>
</tbody>
</table>

54% of pipes have been cloned
RQ1: Characteristics of Yahoo! Pipes Community

<table>
<thead>
<tr>
<th>Metric</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size</td>
<td>8.20 modules per pipe</td>
</tr>
<tr>
<td>Configurability</td>
<td>0.65 modules per pipe</td>
</tr>
<tr>
<td>Popularity</td>
<td>5.67 clones per pipe</td>
</tr>
<tr>
<td>Diversity</td>
<td>3.62 cluster level</td>
</tr>
</tbody>
</table>

Summary

5% of pipes are exact duplicates, yet 46% have a match if field values are relaxed.
RQ1: Characteristics of Yahoo! Pipes Community

Take Aways:
- There is a lot of reuse of shared pipes
- Participants often submit pipes that are highly similar to other pipes in the repository
RQ2: How do pipe attributes change as authors gain experience?

2a: measures experience in terms of time

2b: measures experience in terms of total contributions
RQ2: Analysis of artifacts as authors gain experience

Comparisons based on experience (time)

For each pipe

Get days experience for author

days < 31

yes

add to Early

no

add to Late
RQ2: Analysis of artifacts as authors gain experience

Comparisons based on experience (time)

For each pipe

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>$\mu_{\text{early}}$</th>
<th>$\mu_{\text{late}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Pipes</td>
<td>27,555</td>
<td>5,332</td>
</tr>
<tr>
<td>Diversity***</td>
<td>3.519</td>
<td>4.126</td>
</tr>
<tr>
<td>Popularity***</td>
<td>4.984</td>
<td>9.254</td>
</tr>
<tr>
<td>Configurability***</td>
<td>0.614</td>
<td>0.838</td>
</tr>
<tr>
<td>Size***</td>
<td>7.919</td>
<td>9.587</td>
</tr>
</tbody>
</table>

$H_0: \mu_{\text{early}} > \mu_{\text{late}}$

$H_a: \mu_{\text{early}} \leq \mu_{\text{late}}$

Signif. codes:

*** $0.001$ ** $0.01$
RQ2: Analysis of artifacts as authors gain experience

**Take Away**: More experience results in pipes that are larger, more popular, more configurable, and more diverse
RQ2: Analysis of artifacts as authors gain experience
Comparisons based on contributions

For each author, count all pipes created. If the number of pipes is greater than 15, add pipes to Many; otherwise, add pipes to Few.
RQ2: Analysis of artifacts as authors gain experience
Comparisons based on contributions

For each author, count all pipes created. If the number of pipes is greater than 15, add pipes to Many; otherwise, add pipes to Few.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>$\mu_{\text{few}}$</th>
<th>$\mu_{\text{many}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Pipes</td>
<td>30,503</td>
<td>2,384</td>
</tr>
<tr>
<td>Diversity</td>
<td>3.639</td>
<td>3.355</td>
</tr>
<tr>
<td>Popularity***</td>
<td>4.302</td>
<td>23.250</td>
</tr>
<tr>
<td>Configurability***</td>
<td>0.644</td>
<td>0.729</td>
</tr>
<tr>
<td>Size**</td>
<td>8.194</td>
<td>8.136</td>
</tr>
</tbody>
</table>

H$_0$: $\mu_{\text{few}} > \mu_{\text{many}}$
H$_a$: $\mu_{\text{few}} \leq \mu_{\text{many}}$

Signif. codes:
*** 0.001 ** 0.01
RQ2: Analysis of artifacts as authors gain experience

**Take Away**: The most prolific authors create pipes that are larger, more popular, and more configurable

... what about diversity?
RQ3: What are the characteristics of most prolific authors?

3a: author activities
3b: author skills
3c: awareness of the community
RQ3: Characteristics of most prolific authors

Study Set-up

Authors: 20,313
RQ3: Characteristics of most prolific authors

Study Set-up

Authors: 20,313
Prolific Authors: 81
RQ3: Characteristics of most prolific authors

Rolling Cluster Analysis
RQ3: Characteristics of most prolific authors

Rolling Cluster Analysis
RQ3: Characteristics of most prolific authors

Rolling Cluster Analysis
RQ3: Characteristics of most prolific authors

Rolling Cluster Analysis
RQ3: Characteristics of most prolific authors

Rolling Cluster Analysis
RQ3: Characteristics of most prolific authors

Rolling Cluster Analysis
RQ3: Characteristics of most prolific authors
Rolling Cluster Analysis
RQ3: Characteristics of most prolific authors
Rolling Cluster Analysis
RQ3: Characteristics of most prolific authors
Rolling Cluster Analysis
RQ3: Characteristics of most prolific authors

Rolling Cluster Analysis
RQ3: Characteristics of most prolific authors

Rolling Cluster Analysis
RQ3: Characteristics of most prolific authors

Rolling Cluster Analysis
RQ3: Characteristics of most prolific authors

Rolling Cluster Analysis
RQ3: Characteristics of most prolific authors

Rolling Cluster Analysis
RQ3: Characteristics of most prolific authors

Rolling Cluster Analysis
RQ3: Characteristics of most prolific authors
Rolling Cluster Analysis
RQ3: Characteristics of most prolific authors

Author Activities

Rolling Diversity Analysis Over Time

Time in days: 806 total

Diversity
RQ3: Characteristics of most prolific authors

Author Activities

43% of pipes submitted by prolific authors represent tweaks

For Example: Change a URL, filter criterion, sort order, ...
RQ3: Characteristics of most prolific authors

Author Activities

![Rolling Diversity Analysis Over Time](image)

Level 8: No structural similarities

43% of pipes submitted by prolific authors represent **tweaks**
52% of pipes submitted by prolific authors represent **new initiatives**
RQ3: Characteristics of most prolific authors

Author Activities

Rolling Diversity Analysis Over Time

Time in days:  713  total

56% of prolific authors consistently submit new initiatives
RQ3: Characteristics of most prolific authors

Author Activities

27% of prolific authors consistently submit tweaks
RQ3: Characteristics of most prolific authors

**Take Away #1**: 1/2 of participants submit pipes that are novel to their previous contributions

**Take Away #2**: 1/4 of participants submit pipes that are tweaks of their other pipes
End-user programmer communities may need... 

- moderators.
  - Repository is cluttered with highly similar artifacts (RQ1)
The *real* take away

**End-user programmer communities may need** ...

- moderators.
  - Repository is cluttered with highly similar artifacts (RQ1)
- more sophisticated repository search.
  - Many pipes are very structurally similar to other pipes in the repository (RQ1)
  - Early authors create less diverse pipes than later authors (RQ2)
The *real* take away

*End-user programmer communities may need* . . .

- moderators.
  - Repository is cluttered with highly similar artifacts (RQ1)
- more sophisticated repository search.
  - Many pipes are very structurally similar to other pipes in the repository (RQ1)
  - Early authors create less diverse pipes than later authors (RQ2)
- artifact development support.
  - Tweaks represent missed opportunities for parameterization (RQ3)
  - Many shared pipes are tweaks on previously-committed pipes by the same author (RQ3)
Threats to Validity

- **Internal**
  - History (the pipes were sampled at different times)
  - Selection (the repository only provides public pipes)

- **Construct**
  - Interaction of different factors
  - Mono-method bias on diversity (only consider structural diversity, not semantic)

- **External**
  - Generalizability (only studied one community)
  - Sampling bias (could not control search results when sampling)
Conclusion

- Authors utilize the repository in different ways
- As authors gain experience in the environment, they tend to make more valuable contributions to the repository
- There is a need for better support to help end-user programmer communities continue to progress and grow
- To generalize the results, we are interested in extending the metrics to other languages and repositories

To facilitate replication, the data used in this analysis is available: http://cse.unl.edu/~kstolee/esem2011/artifacts.html