

9-9-2014

When Should I Make Preservation Copies of Myself?

Charles L. Cartledge

Old Dominion University, ccartled@odu.edu

Michael L. Nelson

Old Dominion University, mnelson@odu.edu

Follow this and additional works at: https://digitalcommons.odu.edu/computerscience_presentations



Part of the [Archival Science Commons](#)

Recommended Citation

Cartledge, Charles L. and Nelson, Michael L., "When Should I Make Preservation Copies of Myself?" (2014). *Computer Science Presentations*. 6.

https://digitalcommons.odu.edu/computerscience_presentations/6

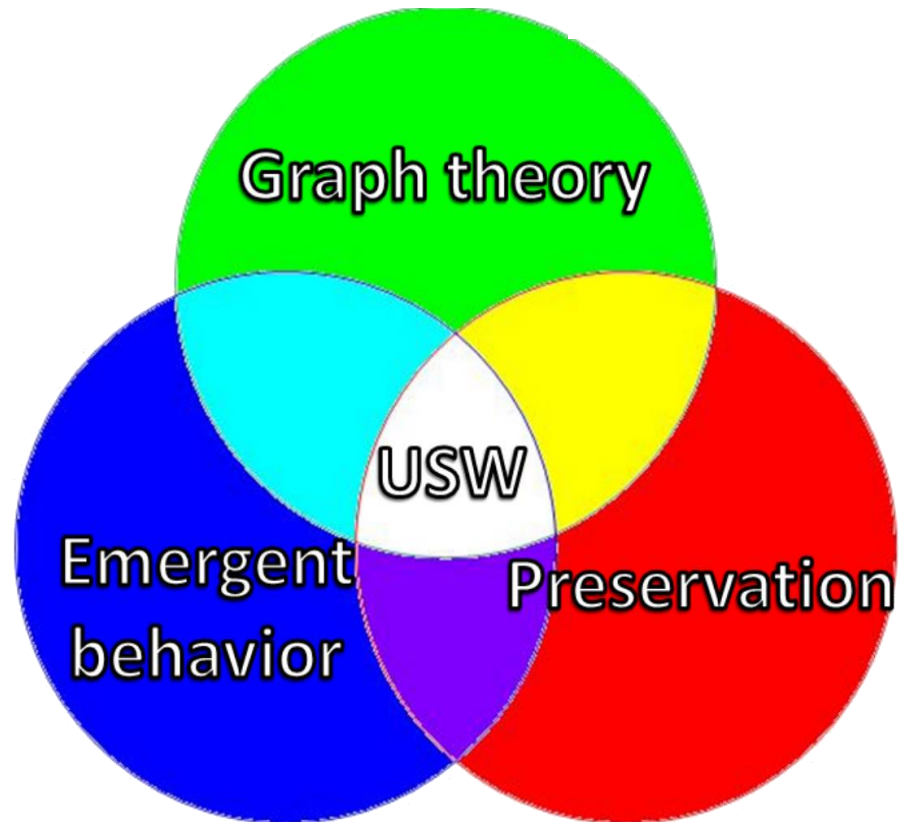
This Book is brought to you for free and open access by the Computer Science at ODU Digital Commons. It has been accepted for inclusion in Computer Science Presentations by an authorized administrator of ODU Digital Commons. For more information, please contact digitalcommons@odu.edu.

When Should I Make Preservation Copies of Myself?

Charles L. Cartledge and Michael L. Nelson
Old Dominion University
Department of Computer Science
Norfolk, VA 23529 USA

JCDL 2014
London, UK
September 9, 2014

Unsupervised Small-World (USW) has multiple areas of interest



Preservation via benign neglect



Handwritten on the back of the photo:

“Josie McClure picture taken Feb 30, 1907 at Poteau, I.T.
Fifteen years of age When this was taken weighed 140 lbs.”

(cultural context needed to make sense of the annotation!)

Will Josie last 100+ years as a web object (WO) in Flickr, Photobucket, et al.?

The image displays two side-by-side browser windows from the Opera browser, illustrating the same vintage portrait of Josie McClure hosted on different platforms.

Left Window (Flickr): The browser address bar shows the URL `www.flickr.com/photos/24791103@N07/6661587389/in`. The Flickr interface includes a "Sign Up" button and navigation links for "Explore" and "Upload". The portrait is centered on a dark background. Below the image, the name "Josie McClure" is displayed, followed by a star icon and a comment icon. A caption at the bottom reads: "Penciled on the back 'Josie McClure picture taken Feb 30, 1907 at Poteau, I.T. Fifteen years of age When this was taken weighed 140 lbs.'"

Right Window (Photobucket): The browser address bar shows the URL `s1370.photobucket.com/user/ccartled1/media/josie_zps70e16cc1.png.html`. The Photobucket interface includes social media sharing icons for Facebook, Twitter, Google+, Pinterest, and Tumblr, along with "Edit" and "Slideshow" buttons. The portrait is centered on a light background. Below the image, the title "Josie McClure, 1907" is displayed, followed by a detailed caption: "An analog artifact in a digital age. Penciled on the back 'Josie McClure picture taken Feb 30, 1907 at Poteau, I.T. Fifteen years of age When this was taken weighed 140 lbs.'". A "Tags" field with an "Add tag" button is visible below the caption. A small thumbnail of the image is shown at the bottom center. In the bottom right corner, there are icons for views (9), comments (0), and likes (0).

Crowd sourcing preservation

- “Everyone is a curator ...”
 - Crowd sourced activity
 - Unscheduled
 - Willing to wait a long time
- Enlist humans in creation and maintenance – opposite of benign neglect

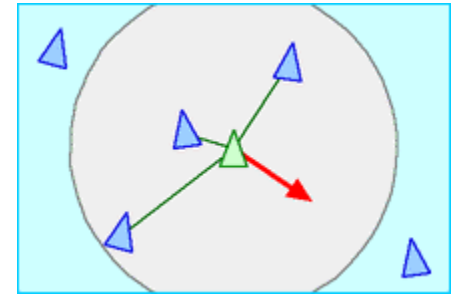


Frank McCown, Michael L. Nelson, and Herbert Van de Sompel, Everyone is a Curator: Human-Assisted Preservation for ORE Aggregations, Proceedings of the DigCCurr 2009 <http://arxiv.org/abs/0901.4571>

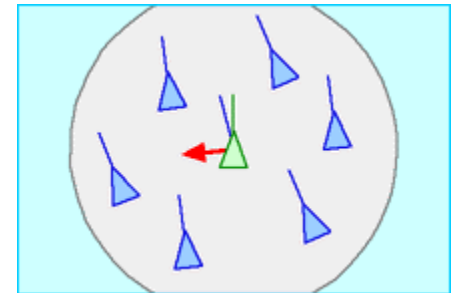
See also: <http://ws-dl.blogspot.com/2013/10/2013-10-23-preserve-me-if-you-can-using.html>

Emergent behavior: flocking boids

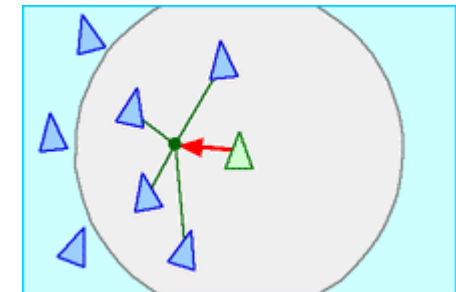
- Craig Reynolds – basis of herd and flock behavior in computer animations
 - 3 rules
 - Collision avoidance
 - Velocity matching
 - Flock centering
 - No central control, everything based on local knowledge only
- **Simple rules produce complex, emergent behavior**



Collision avoidance



Velocity matching



Flock centering

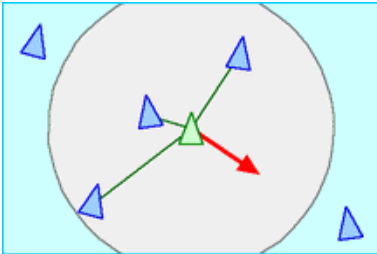
Craig W. Reynolds, Computer Animation with Scripts and Actors, ACM SIGGRAPH, vol. 16, ACM, 1982, pp. 289 - 296.

Images <http://www.red3d.com/cwr/boids/>

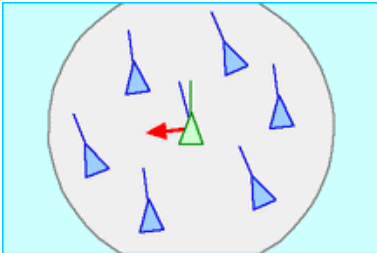
USW interpretation of flocking

Craig Reynolds' "boids"

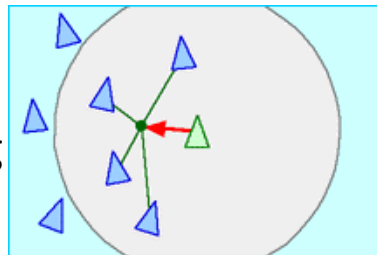
Collision avoidance



Velocity matching



Flock centering



USW interpretation

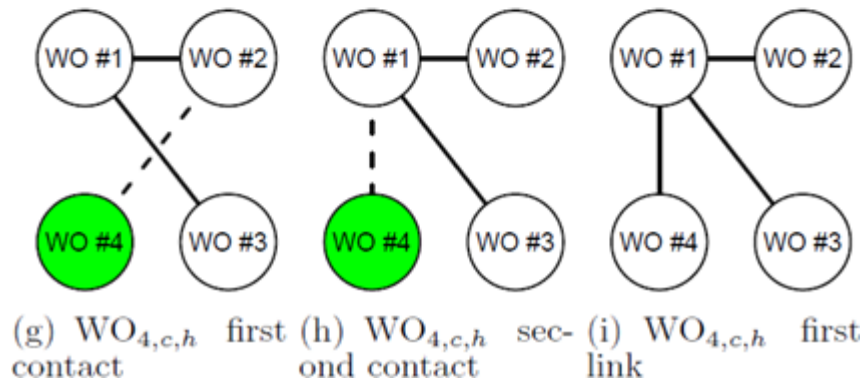
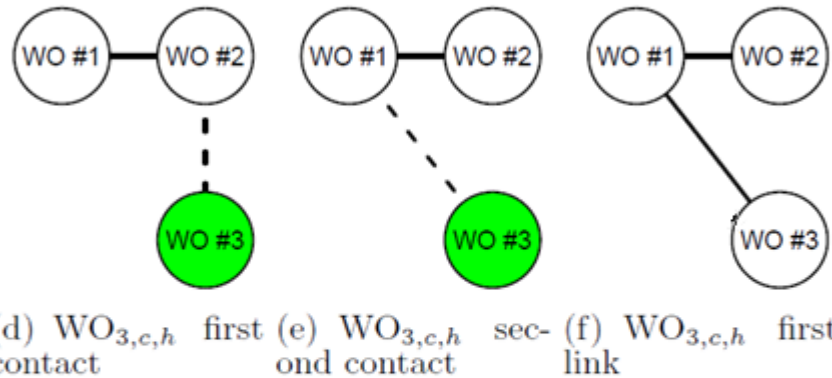
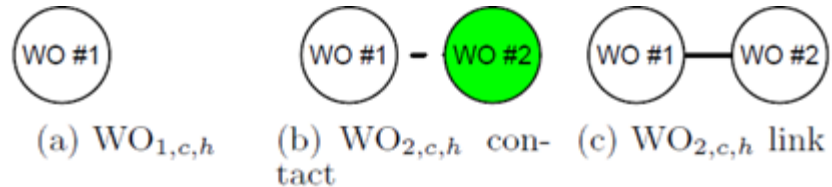
Each WO has a unique URI

Matching number of copies/family members

Move with friends to new hosts

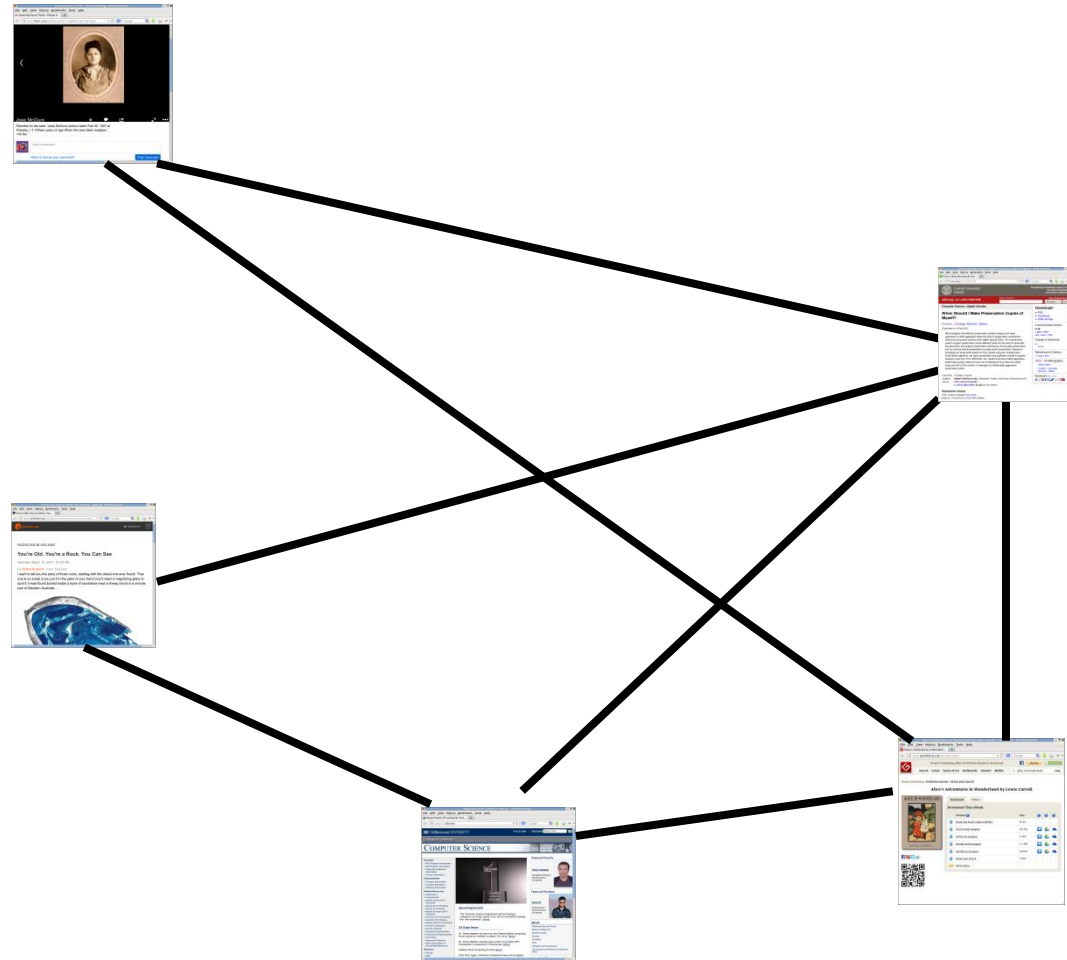
WOs wandering in the USW graph

- Wandering WO is “introduced” to an existing WO
- If a connection is not made, then an attempt is made to another existing WO
- Process is repeated until a connection is made
- **No global knowledge**
 - No omnipotent enforcer
 - No omnipresent monitor
- **No repositories**



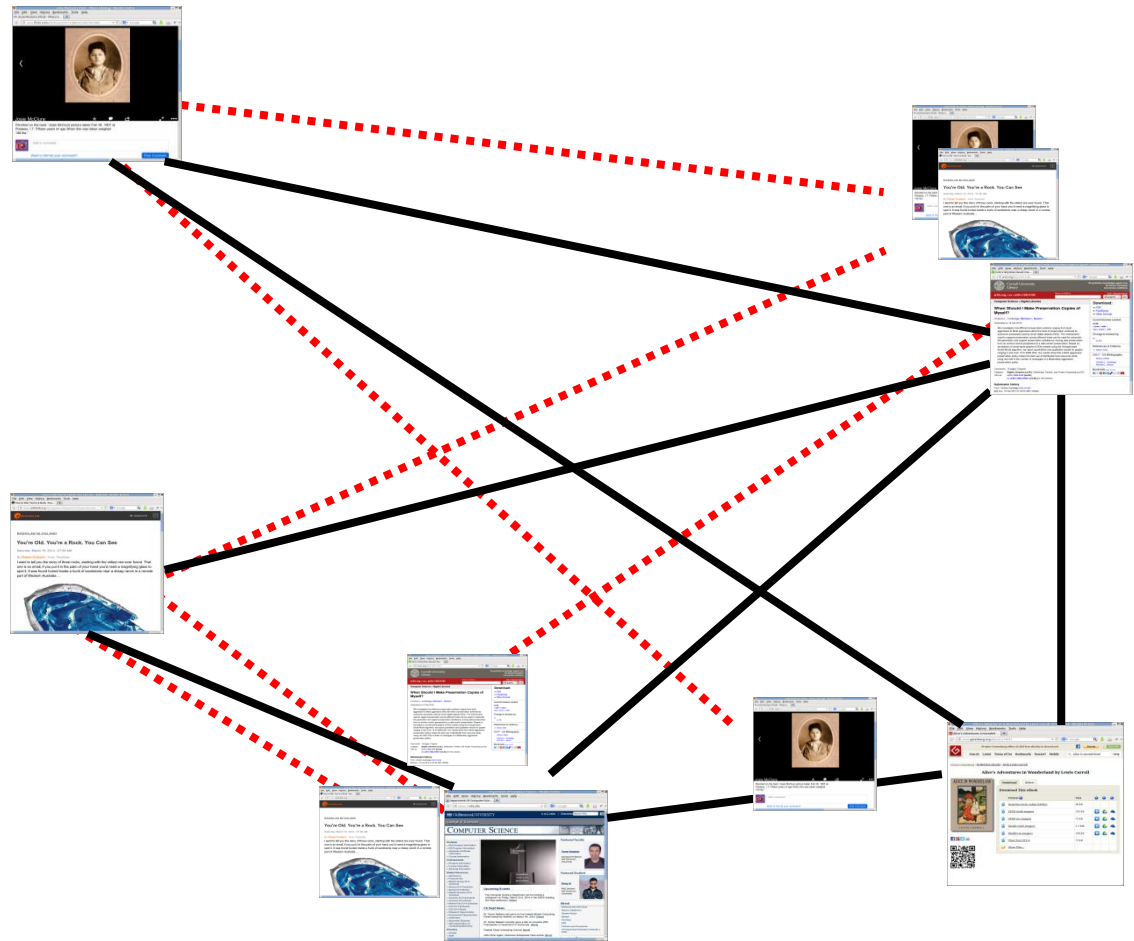
USW WO “friendship” links

- WOs have “friendship” links to other WOs
- Different than HTML navigational links (i.e., `<link>` instead of `<a>`)



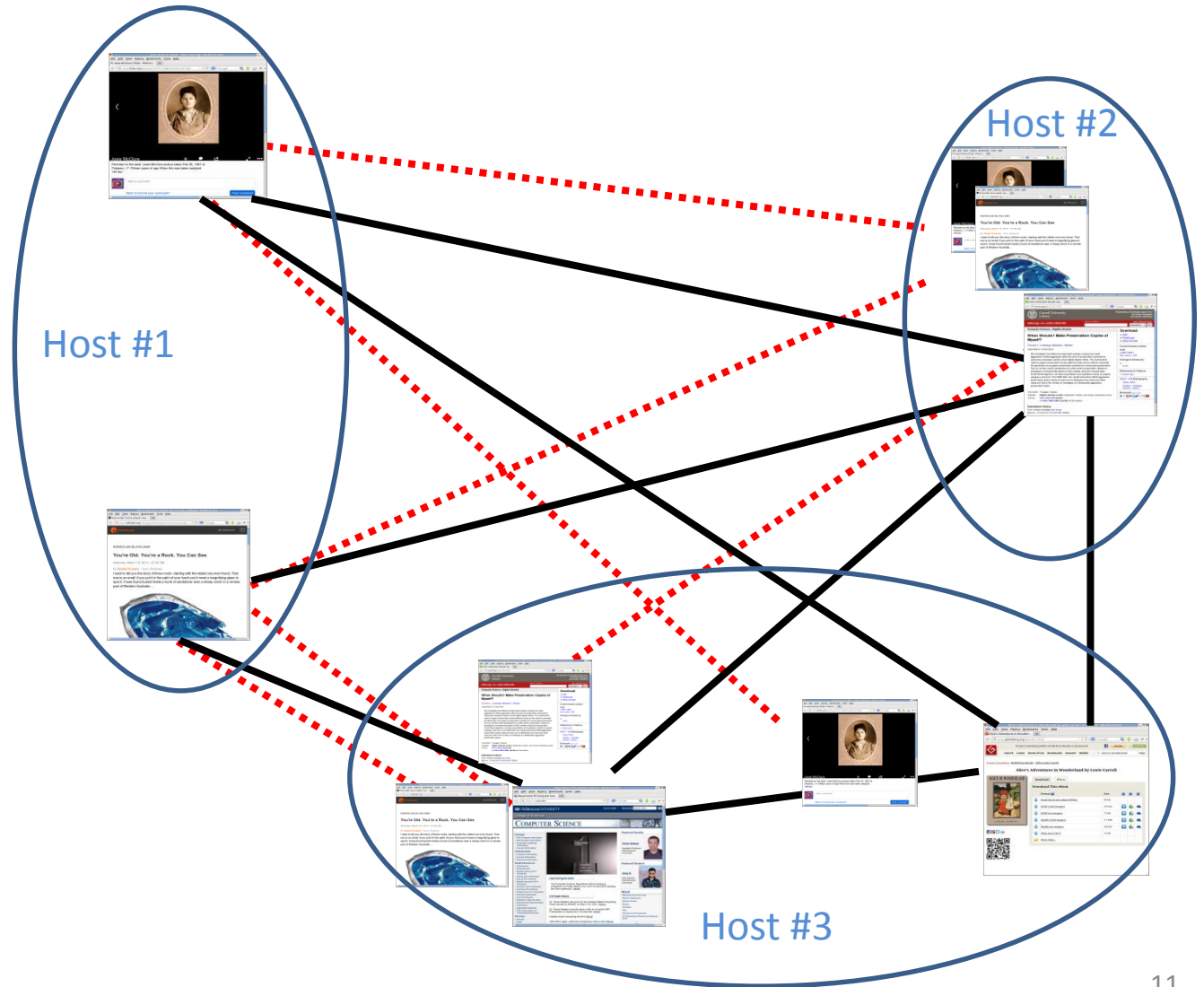
USW WO “families”

A family is a set of copies of the same WO



USW hosts

Family members live on different hosts



WO roles & responsibilities

- Hierarchy of family WOs
 - Progenitor – initial WO
 - Copies – more recent WO copies
 - Each WO is timestamped with creation time
- WO roles
 - Active maintainer – eldest WO charged with making copies and related housekeeping
 - Passive maintainer – all other WOs
- Order of precedence
 - If progenitor is accessible then it is the active maintainer
 - If declared active maintainer is accessible then it is the active maintainer
 - Otherwise, WO declares itself active maintainer
- If family is disconnected then multiple active maintainers are possible until reconnection then the **eldest** WO declares itself active maintainer

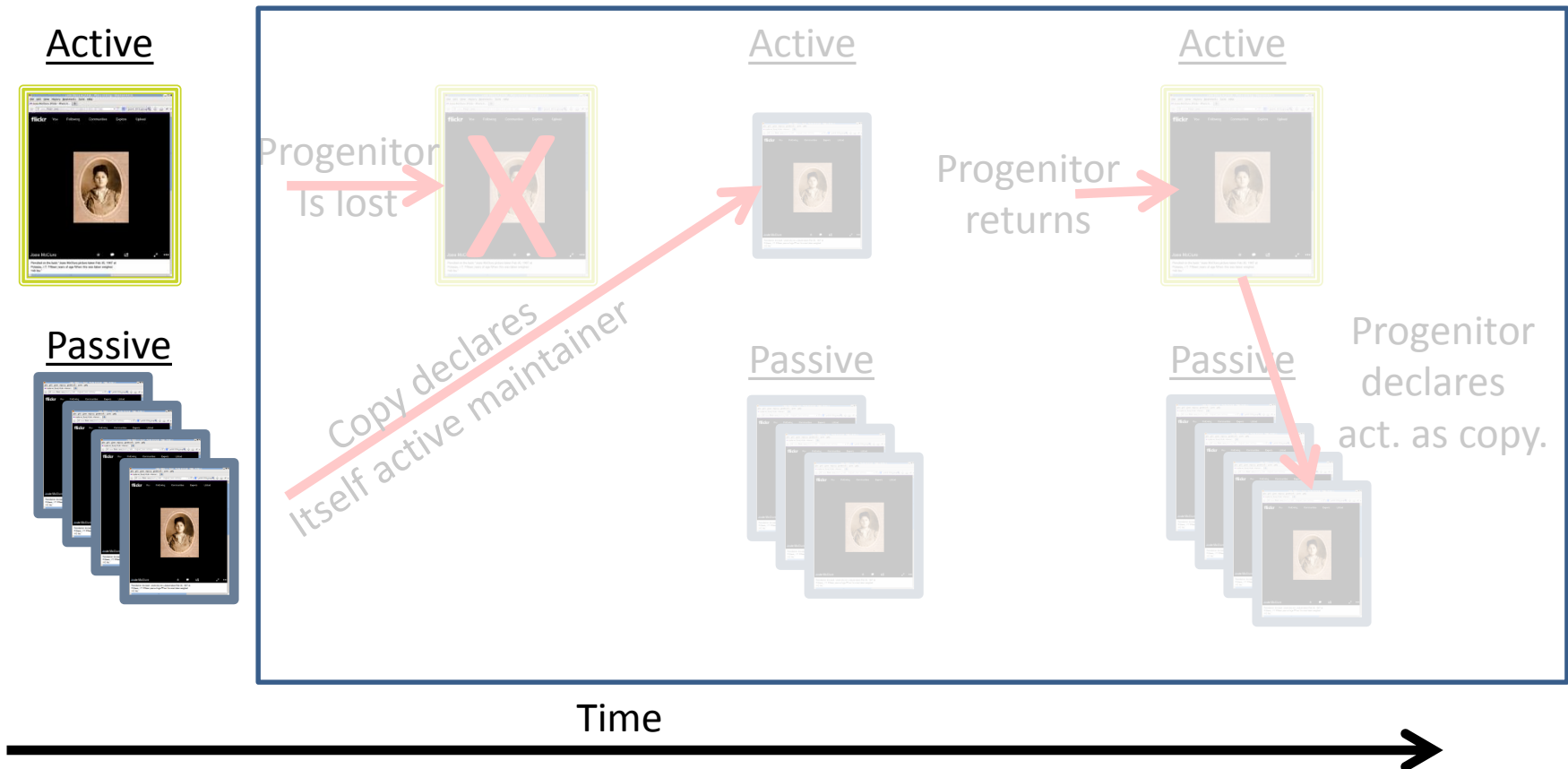


Progenitor



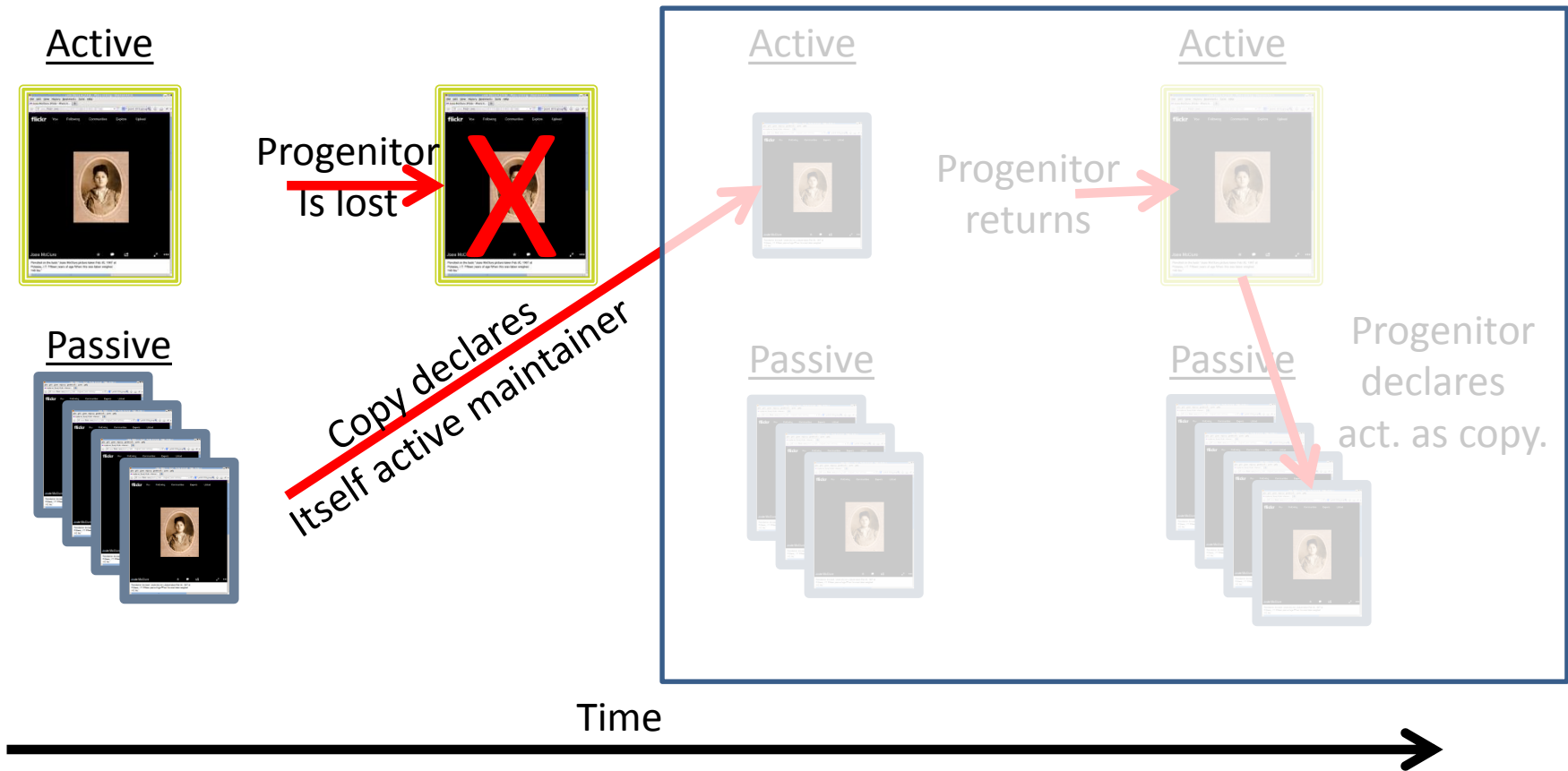
Copies

Active and passive maintenance activities



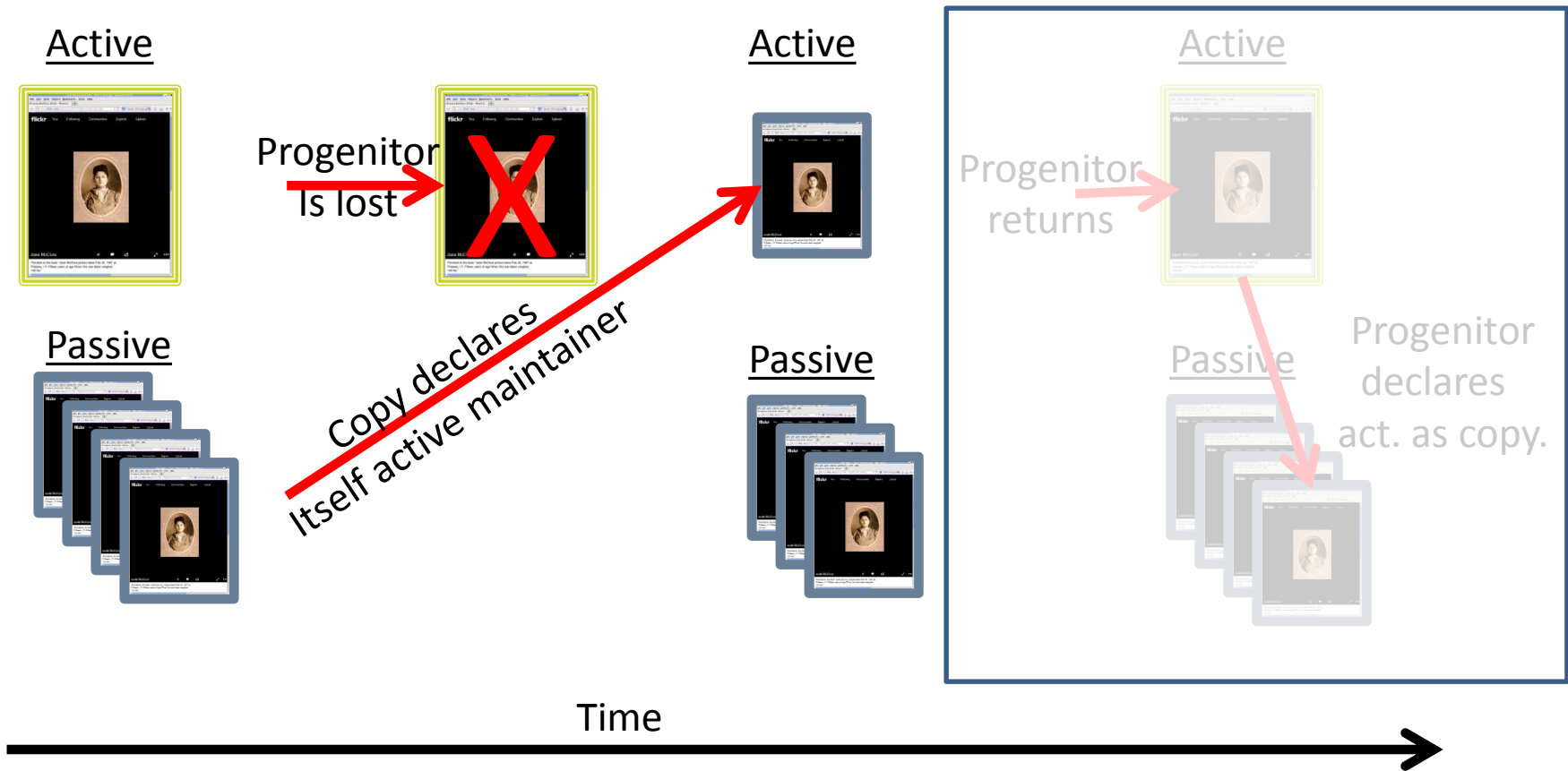
- Active maintainer (the WO with earliest timestamp) – currently charged with making copies and related housekeeping
- Passive maintainer – all other WOs

Progenitor is lost



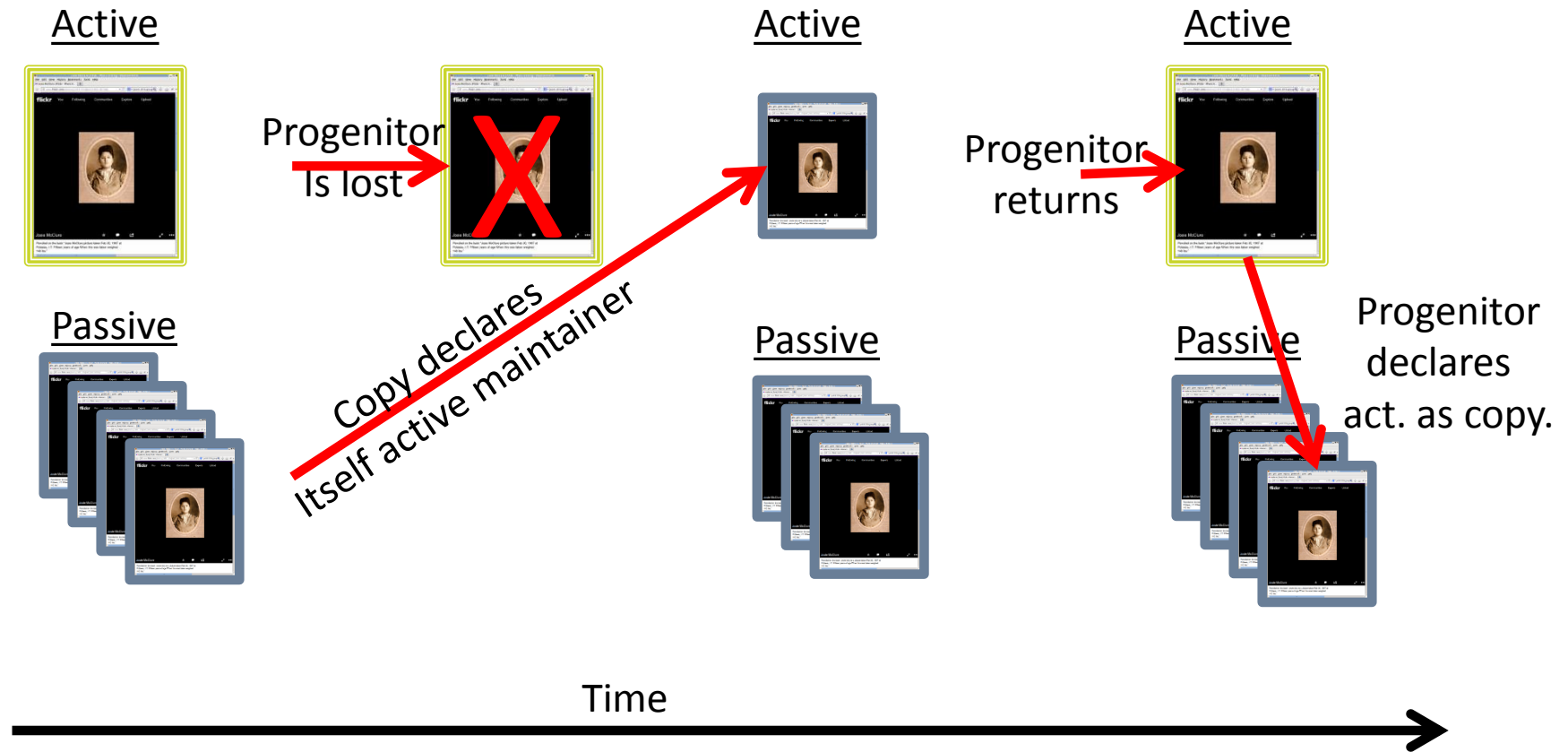
- Active maintainer – currently charged with making copies and related housekeeping
- Passive maintainer – all other WOs

A new active maintainer



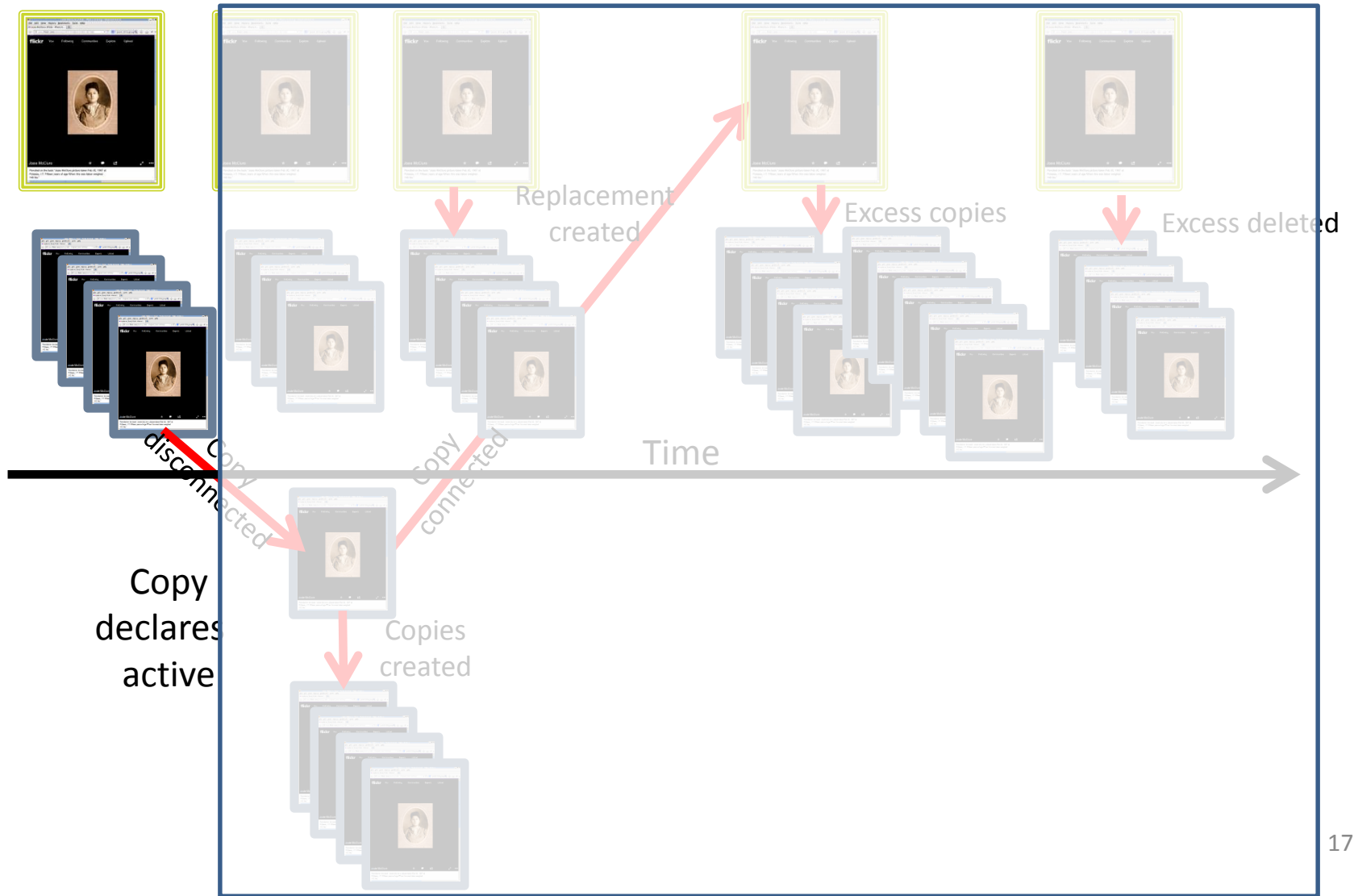
- Active maintainer – currently charged with making copies and related housekeeping
- Passive maintainer – all other WOs

Progenitor returns and assumes active maintainer role

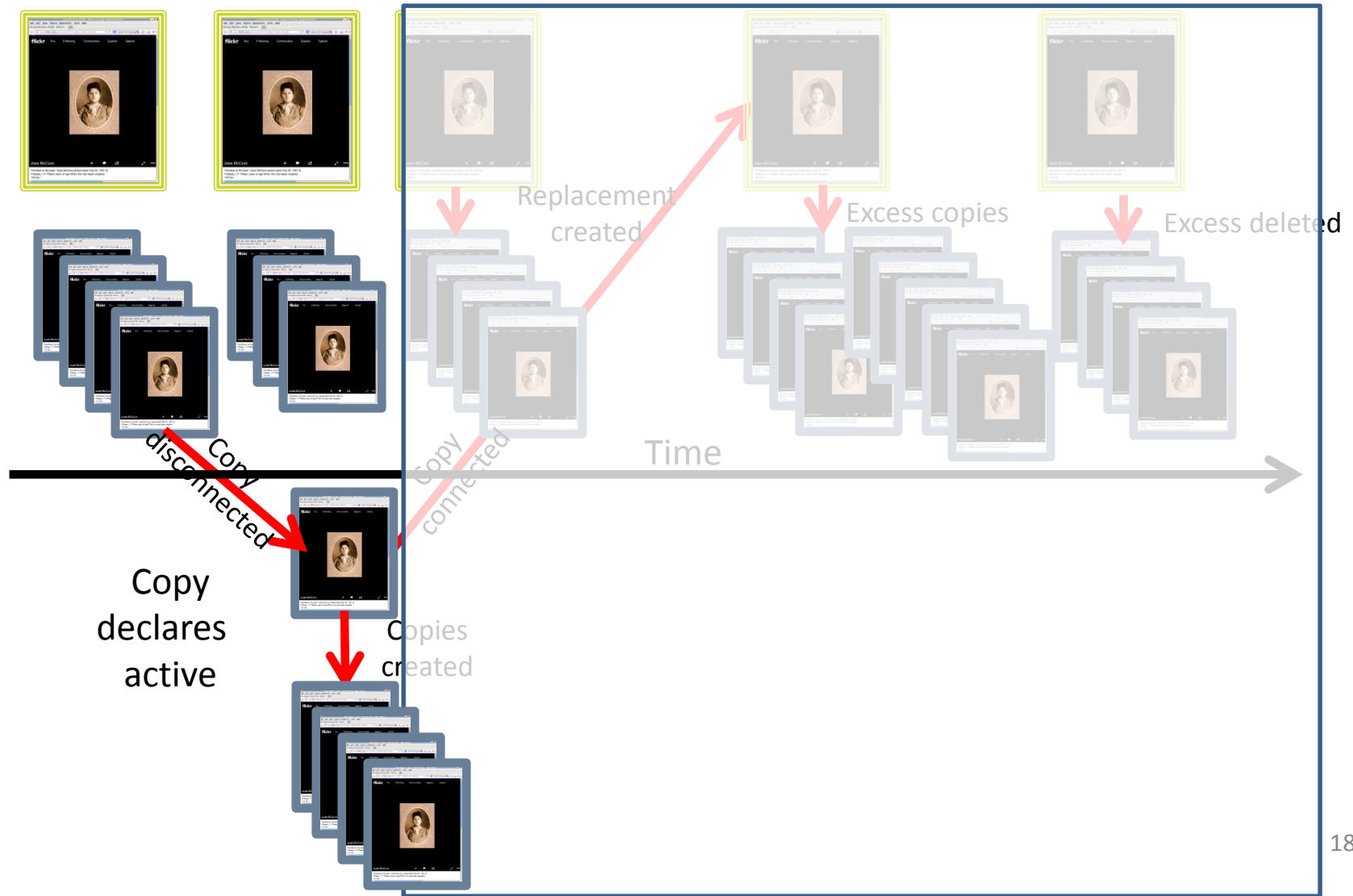


- Active maintainer – currently charged with making copies and related housekeeping
- Passive maintainer – all other WOs

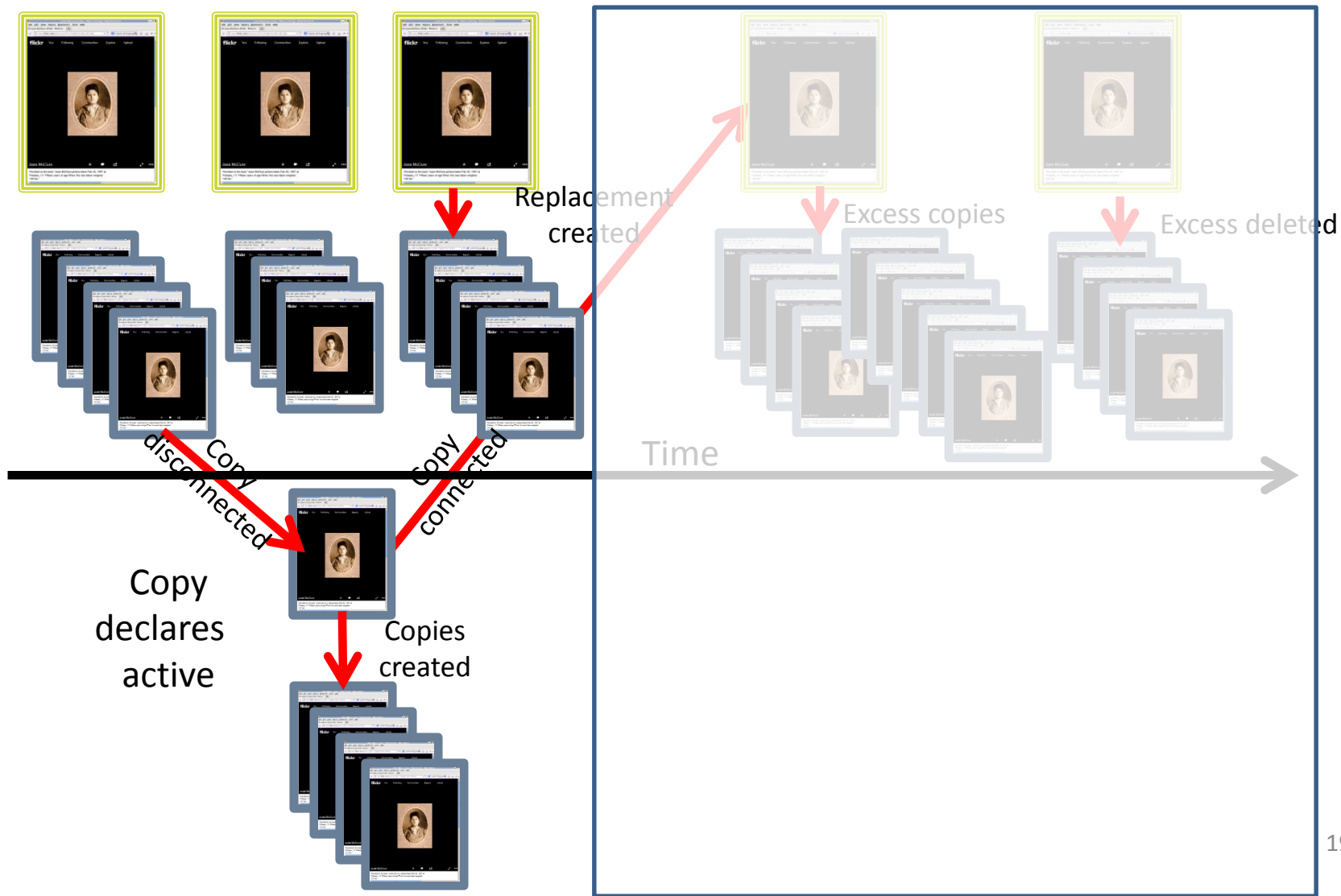
Progenitor has made copies



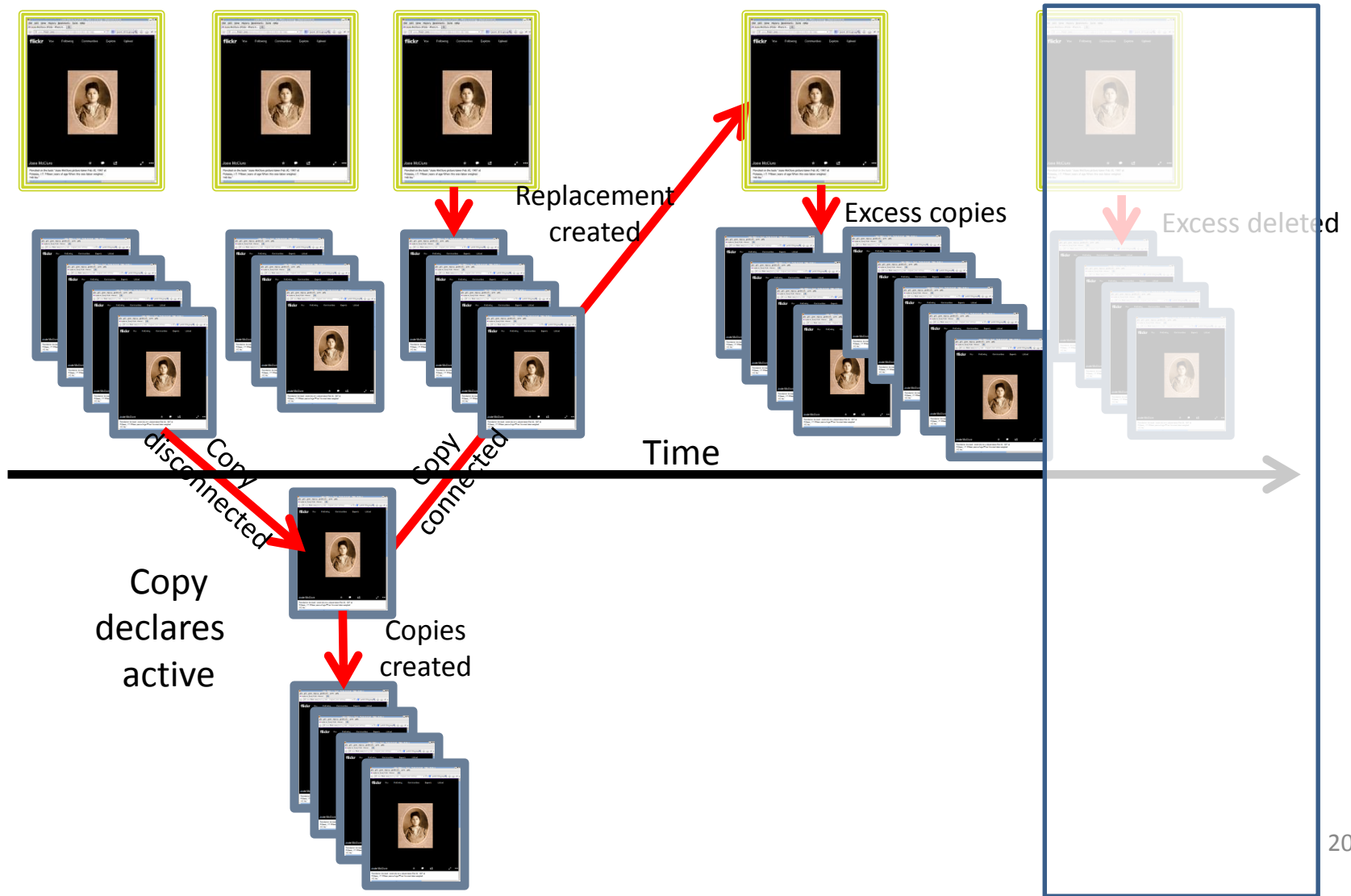
A copy is disconnected from the family



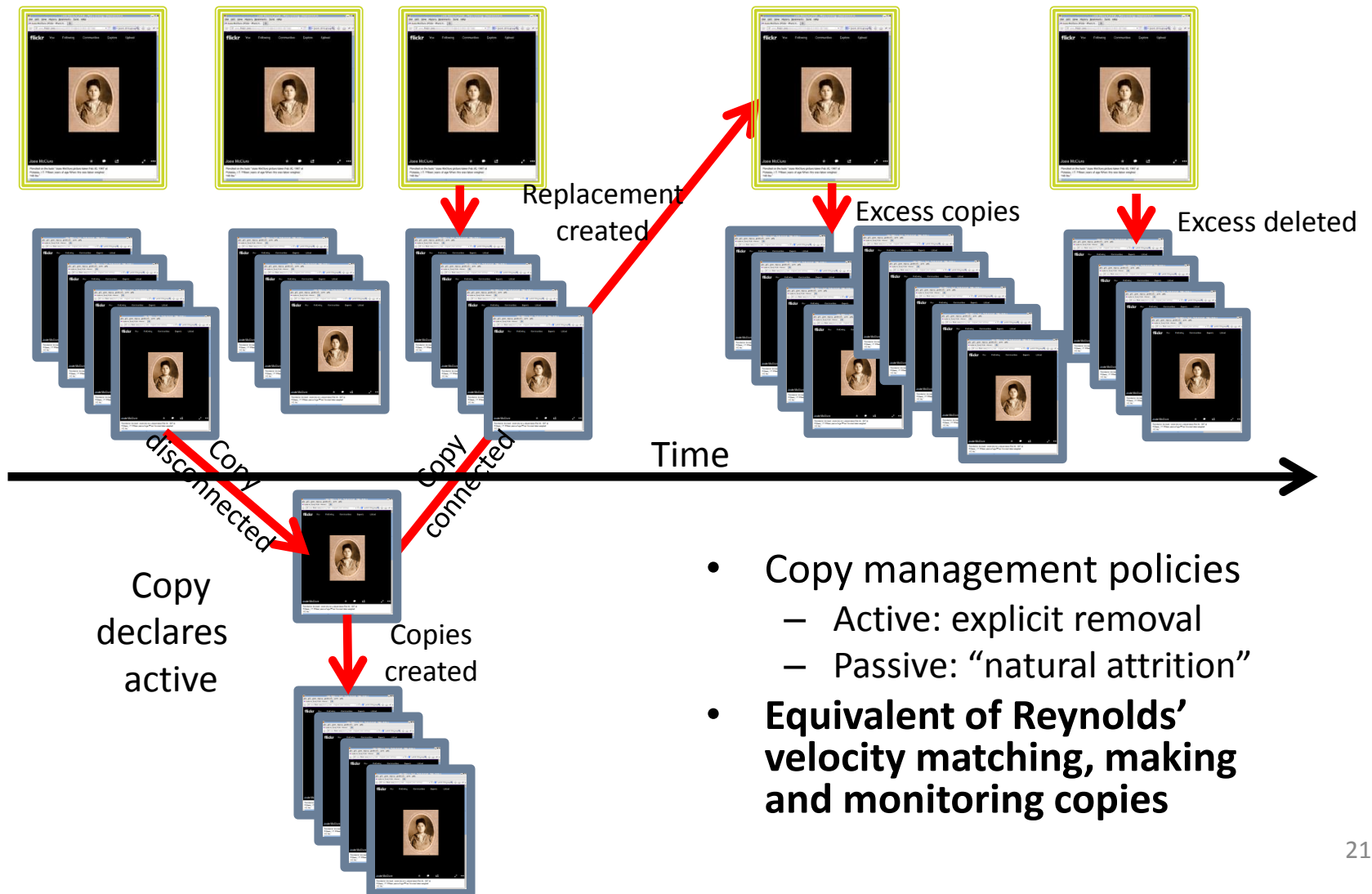
Two active maintainers make copies



Disconnected copy is reconnected to the progenitor



Family has too many copies



- Copy management policies
 - Active: explicit removal
 - Passive: “natural attrition”
- **Equivalent of Reynolds’ velocity matching, making and monitoring copies**

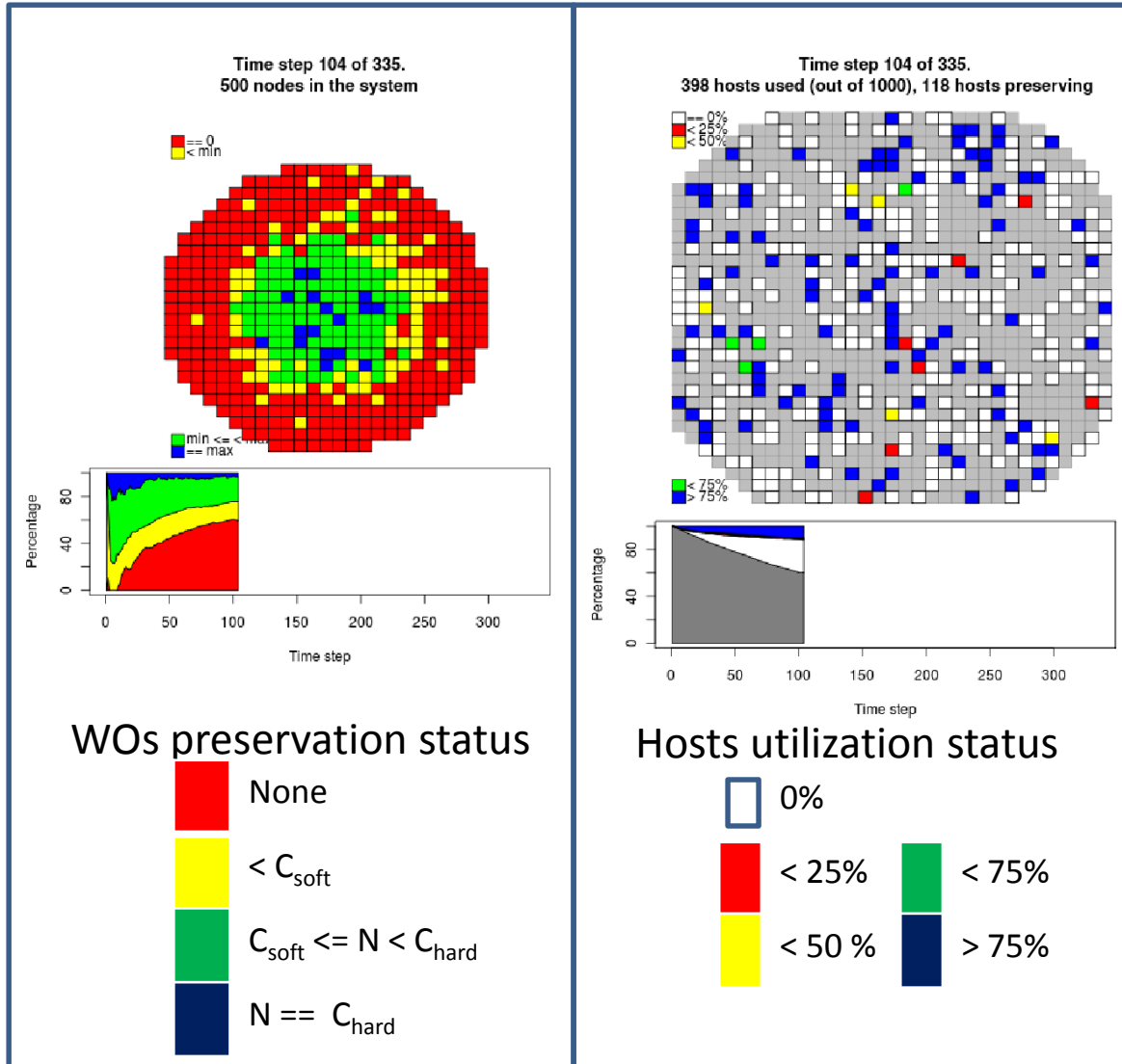
Parameters

- c_{soft} = minimum number of preservation copies desired by a web object
 - e.g., $c_{\text{soft}} = 3$
- c_{hard} = maximum number of preservation copies desired by a web object
 - e.g., $c_{\text{hard}} = 5$
- h_{max} = maximum number of hosts
 - e.g., $h_{\text{max}} = 1000$
- h_{cap} = host capacity for web objects
 - e.g., $h_{\text{cap}} = 5$
- n_{max} = maximum number of web objects
 - e.g., $n_{\text{max}} = 500$

Three USW copying policies

- **Least aggressive** – one at a time to c_{hard}
- **Moderately aggressive** – as quickly as possible to c_{soft} and then one at a time c_{hard}
- **Most aggressive** – as quickly as possible to c_{hard}
- **Constraints:**
 - WOs can only take action when woken up by interactive users or other WOs (i.e., mostly they lie dormant waiting for crowd sourced preservation)
 - Copying continues until WOs can no longer find hosts that are not full

Reading tree ring graphs



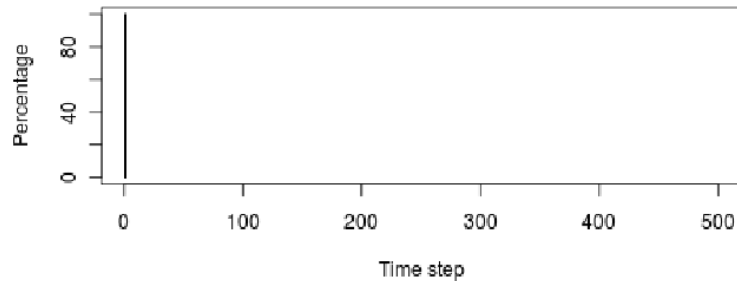
Least aggressive (t = 1)

Time step 1 of 334.
5 nodes in the system

■ == 0
■ < min



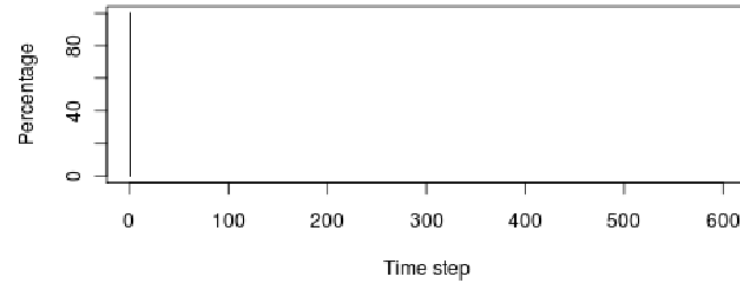
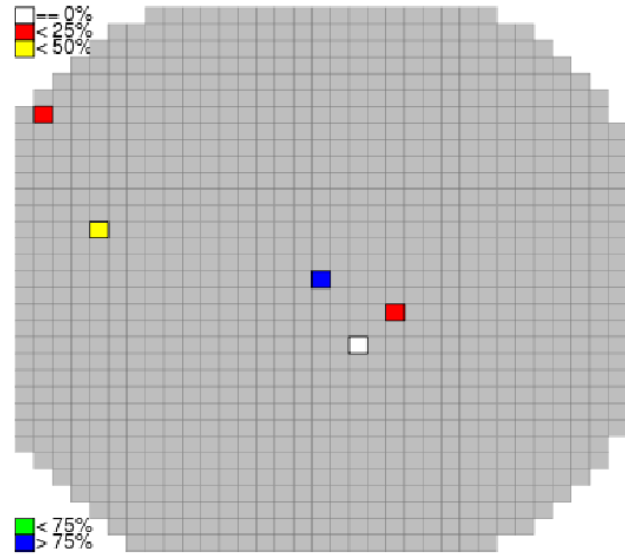
■ min <= < max
■ == max



Time step 1 of 334.
5 hosts used (out of 1000), 4 hosts preserving

■ == 0%
■ < 25%
■ < 50%

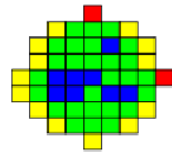
■ < 75%
■ < 75%



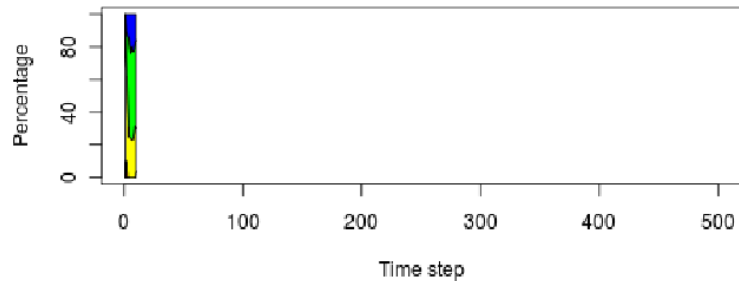
Least aggressive (t = 10)

Time step 10 of 334.
50 nodes in the system

■ == 0
■ < min

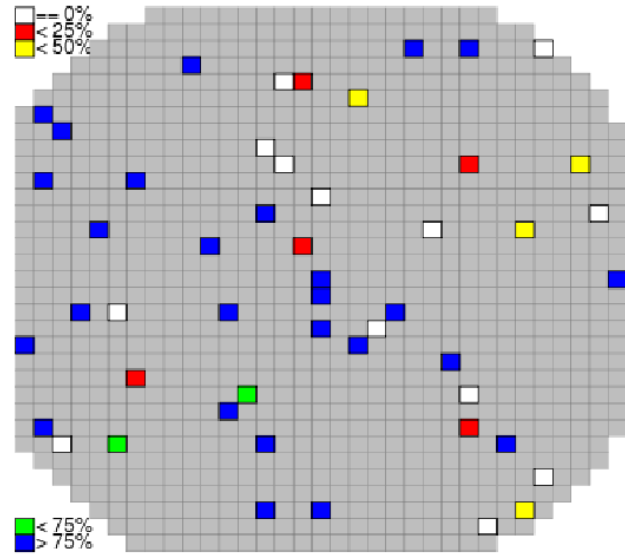


■ min <= < max
■ == max

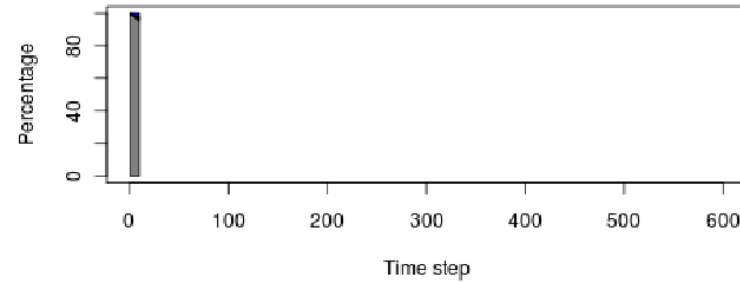


Time step 10 of 334.
50 hosts used (out of 1000), 37 hosts preserving

■ == 0%
■ < 25%
■ < 50%



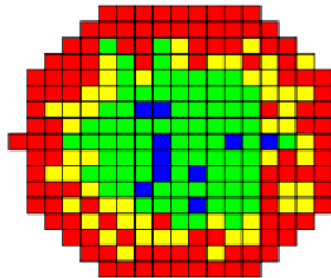
■ < 75%
■ > 75%



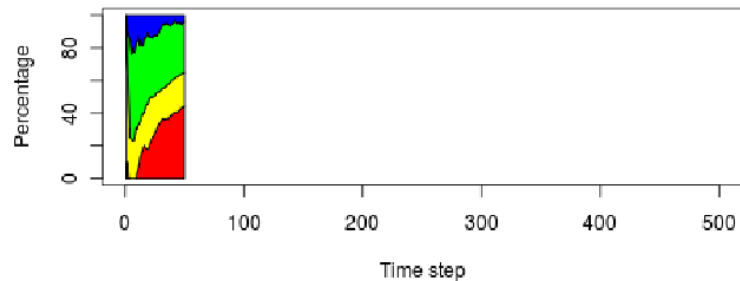
Least aggressive (t = 50)

Time step 50 of 334.
250 nodes in the system

== 0
< min

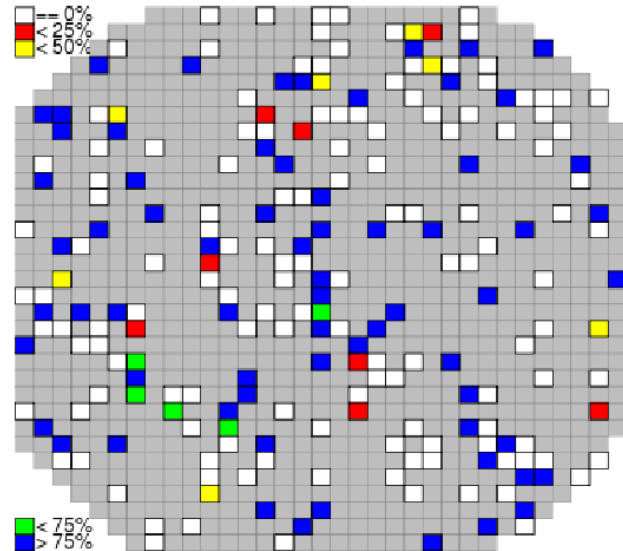


min <= < max
== max

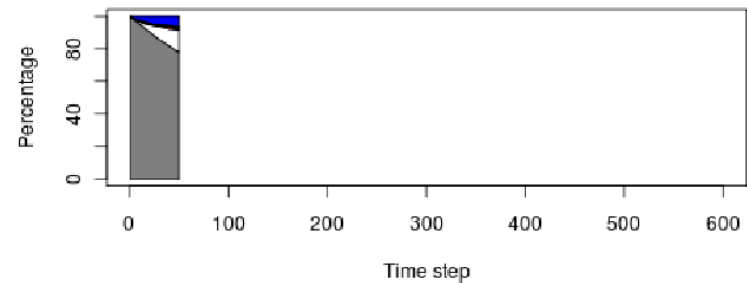


Time step 50 of 334.
222 hosts used (out of 1000), 87 hosts preserving

== 0%
< 25%
< 50%

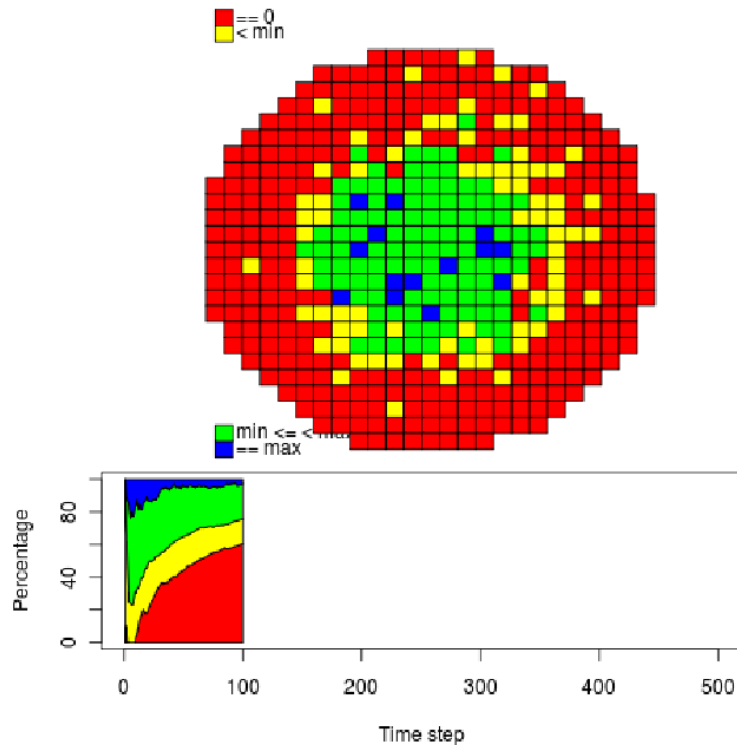


> 75%
> 75%

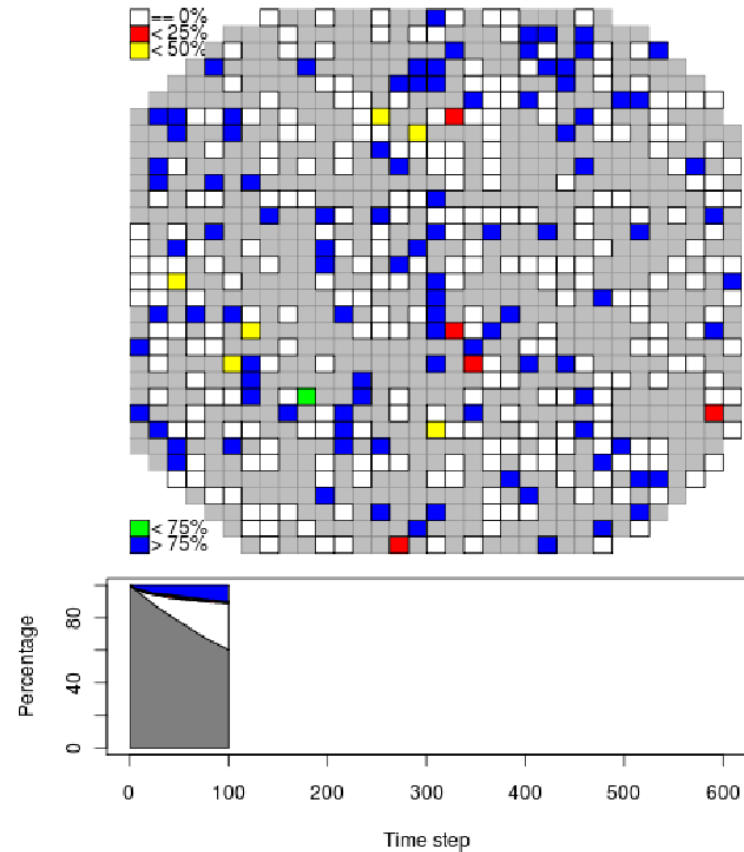


Least aggressive (t = 100)

Time step 100 of 334.
500 nodes in the system



Time step 100 of 334.
398 hosts used (out of 1000), 113 hosts preserving

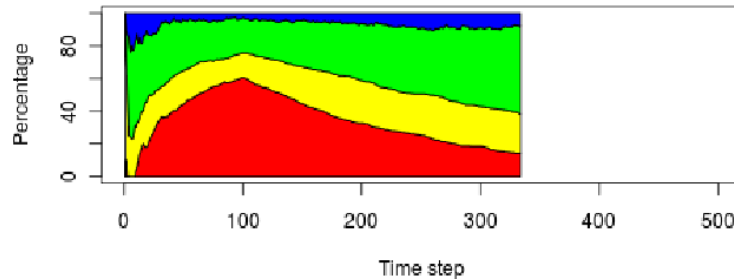
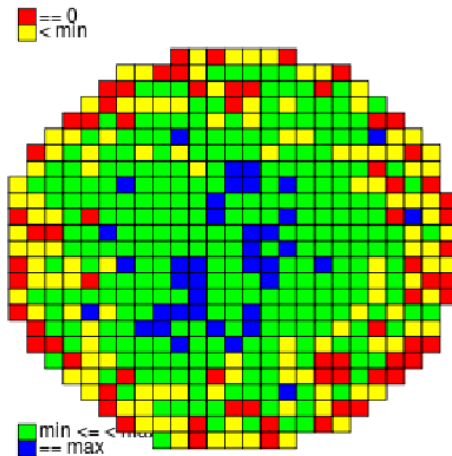


□ A full YouTube video is available at: <http://youtu.be/sHJGYphqtK4>

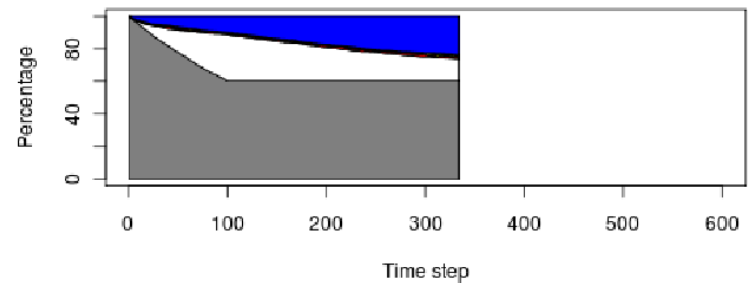
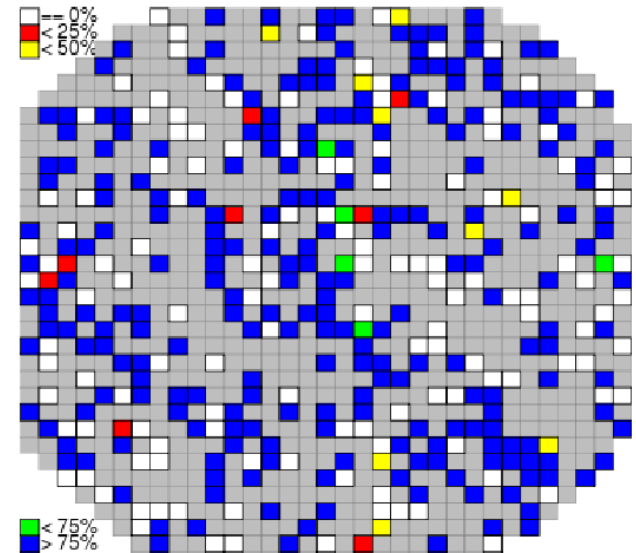
Least aggressive (final)

- Results
 - System stabilized
 - Host capacity limited
 - Some WOs without any copies
 - Some hosts unused
- “Least aggressive” is not an effective policy

Time step 334 of 334.
500 nodes in the system

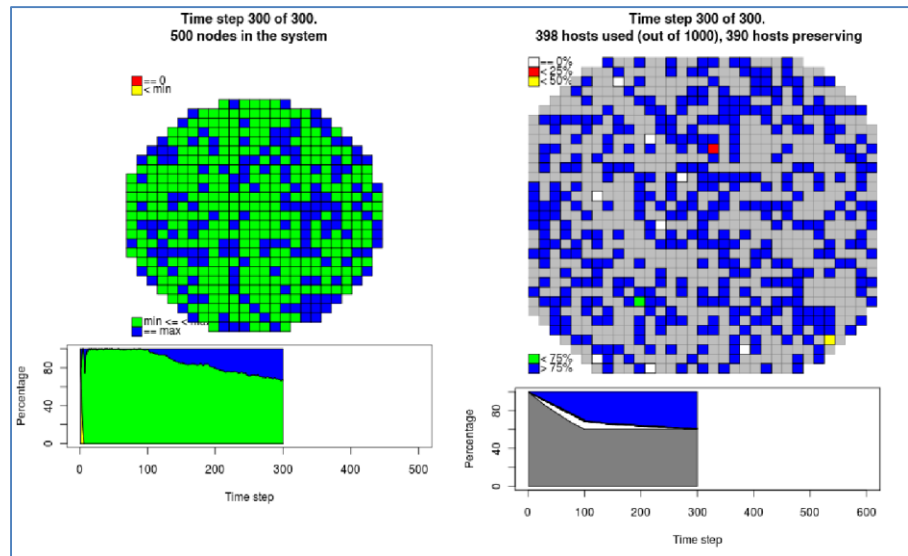
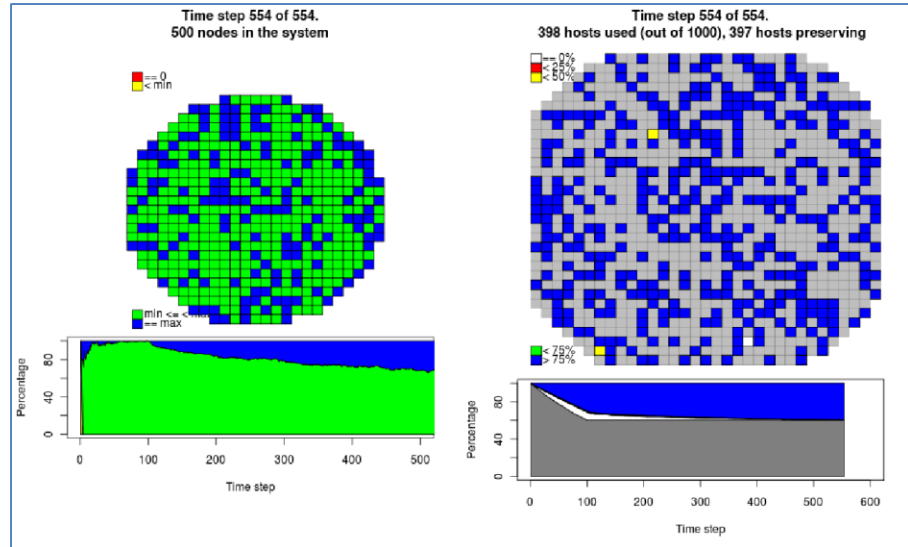


Time step 334 of 334.
398 hosts used (out of 1000), 263 hosts preserving

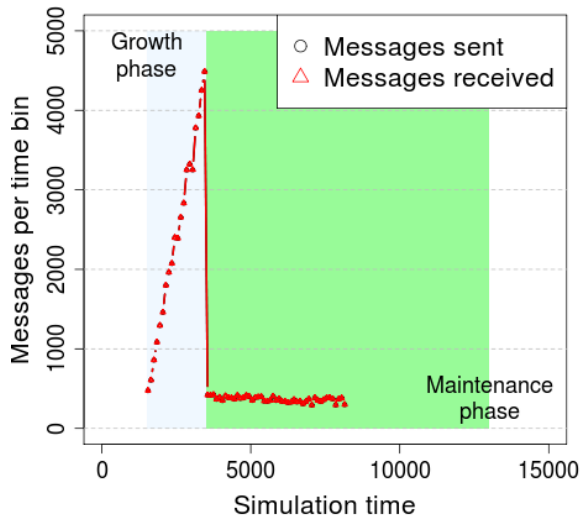


Which policy to choose?

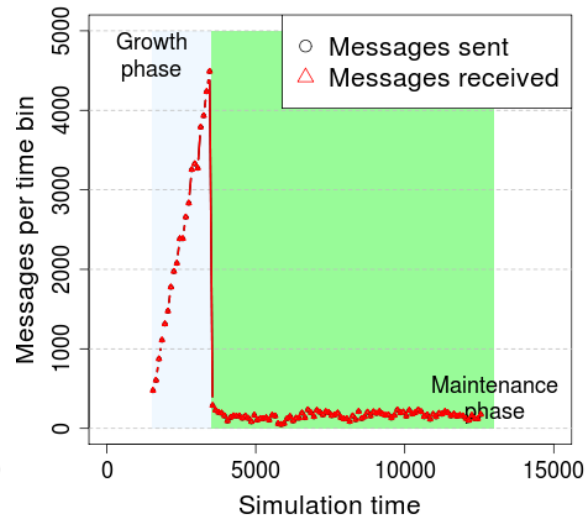
- **Moderately aggressive** results in an additional 18% of WOs meeting their preservation goals and makes more efficient use of limited host resources sooner
- **Most aggressive** results in almost the same percentage of WOs meeting their goals, but with slightly more hosts having unused capacity



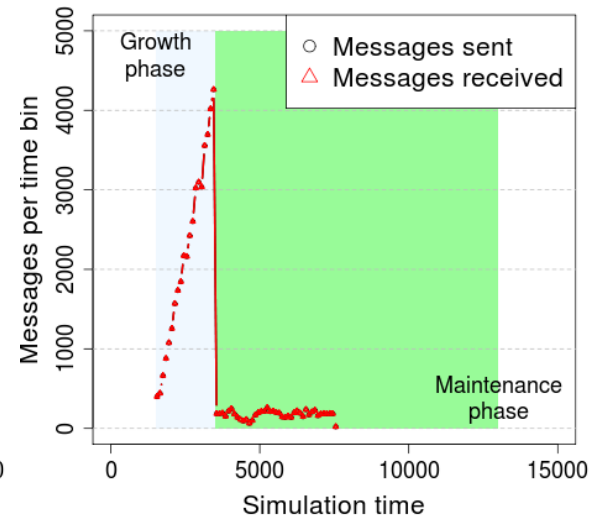
How does policy affect message exchange?



Least aggressive



Moderately aggressive



Most aggressive

Number of messages is constant, but amortized over different time scales

Conclusions

- Based on simulations:
 - Be aggressive when making copies!
 - Moderately aggressive copying was approximately the same as aggressive copying
 - Aggressive achieves steady state faster
 - But moderately aggressive distributes WOs over hosts more equally
 - Moderately aggressive vs. aggressive comes down to “go fast” vs. “spread the load”

Video URLs

- USW video
 - <http://youtu.be/JnCMenp73YQ>
- Least Aggressive
 - <http://youtu.be/sHJGYphqtK4>
- Moderately Aggressive
 - <https://www.youtube.com/watch?v=pVI-VhPh7KQ>
- Most Aggressive
 - <https://www.youtube.com/watch?v=eIXz8Njh-QM>
- “Death Star” message histogram
 - <https://www.youtube.com/watch?v=X3EShyjFoc4>
- “Traditional” message histogram
 - <https://www.youtube.com/watch?v=9CcCup3Td-Q>

Backup slides

Some WO reference implementation details

Direct WO to WO communication:
simulated via the HTTP Mailbox

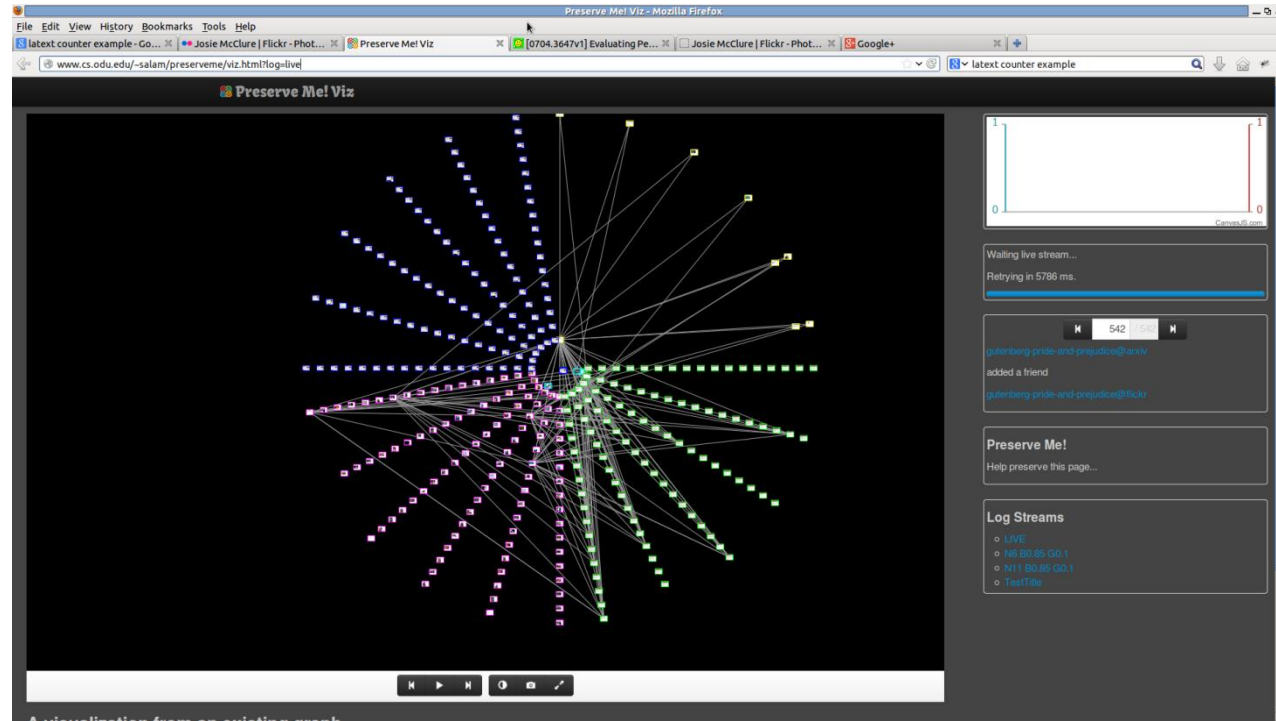
WO memory: simulated via “edit”
service

```
<link rel="alternate" type="text/html" href="http://flickr.cs.odu.edu/flickr-24/91" />
<link rel="self" type="application/atom+xml" href="http://flickr.cs.odu.edu/remsf" />
<link rel="edit" type="application/atom+xml" href="http://ws-dl-02.cs.odu.edu:1010" />
<link rel="http://wsdl.cs.odu.edu/uswdo/terms/copy" type="application/atom+xml" href="http://wsdl.cs.odu.edu/uswdo/terms/copy" />
<link rel="http://wsdl.cs.odu.edu/uswdo/terms/synchronize" type="message/http" href="http://wsdl.cs.odu.edu/uswdo/terms/synchronize" />
<link rel="http://wsdl.cs.odu.edu/uswdo/terms/httpmailbox#self" href="http://ws-dl-02.cs.odu.edu/uswdo/terms/httpmailbox#self" />
<link rel="http://wsdl.cs.odu.edu/uswdo/terms/httpmailbox#all" href="http://ws-dl-02.cs.odu.edu/uswdo/terms/httpmailbox#all" />
<link rel="http://wsdl.cs.odu.edu/uswdo/terms/httpmailbox#family" href="http://ws-dl-02.cs.odu.edu/uswdo/terms/httpmailbox#family" />
<link rel="http://www.openarchives.org/ore/terms/describes" href="http://flickr.cs.odu.edu/flickr-24/91" />
```

- ❑ Sawood Alam, HTTP Mailbox - Asynchronous RESTful Communication, Master's thesis, Old Dominion University, Norfolk, VA, 2013.
- ❑ Carl Lagoze, Herbert Van de Sompel, Pete Johnston, Michael Nelson, Robert Sanderson, and Simeon Warner, ORE User Guide - Resource Map Implementation in Atom, Tech. report, Open Archives Initiative, 2004.
- ❑ Sawood Alam, Charles L. Cartledge, and Michael L. Nelson, Support for Various HTTP Methods on the Web, Tech. Report arXiv:1405.2330 (2014).

Preserve Me Viz! with new connections

- New friend connections
- New copy locations



Preserve Me “Basic” on a copy

- Differences between active and passive maintainers.
- Active maintainer is responsible for making copies.
- Passive maintainer sends alerts to the active maintainer
- Passive maintainer may assume active maintainer role if active is not available.

Preserve Me! - Mozilla Firefox

www.cs.odu.edu/~salam/preserveme/preserveme.html?rem=http%3A%2F%2Fflickr.cs.odu.edu%2Fcopyrems%2Fflickr-24791103-N07-12867674403.xml

Preserve Me! Advanced

courtDeTomasDeTorquemada

Author(s)
courtDeTomasDeTorquemada
simple_simon_2007

Status: Safe! (Number of copies: 1) Updated: 27 minutes ago

Check Health!

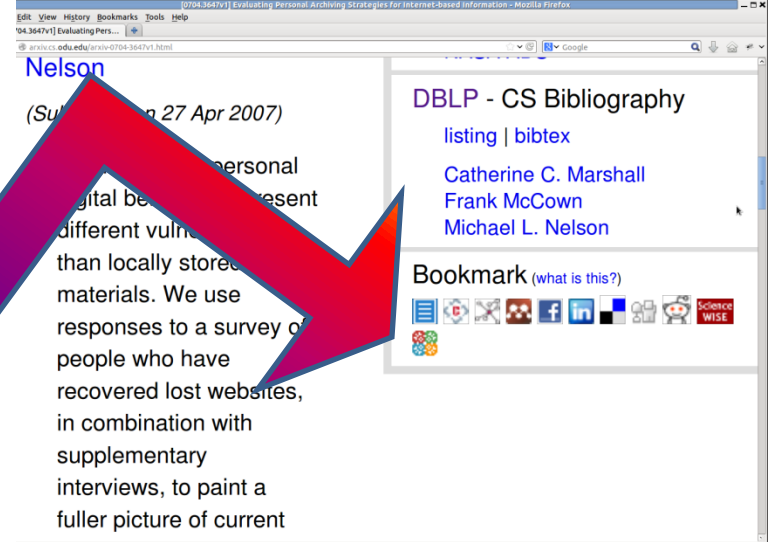
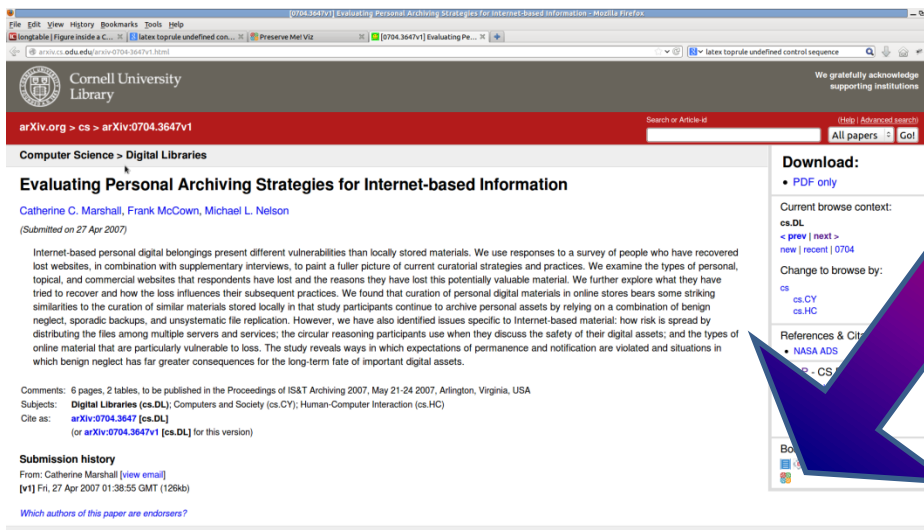
This Resource Map was not updated recently. Check to see if there are any pending maintenance tasks that will change the status of the Resource Map. [Need Help?](#)

Check Health

Existing Copies

● <http://flickr.cs.odu.edu/rems/flickr-24791103-N07-12867674403.xml> Original

A USW instrumented splash page



Nelson
(Submitted on 27 Apr 2007)
Internet-based personal digital belongings present different vulnerabilities than locally stored materials. We use responses to a survey of people who have recovered lost websites, in combination with supplementary interviews, to paint a fuller picture of current

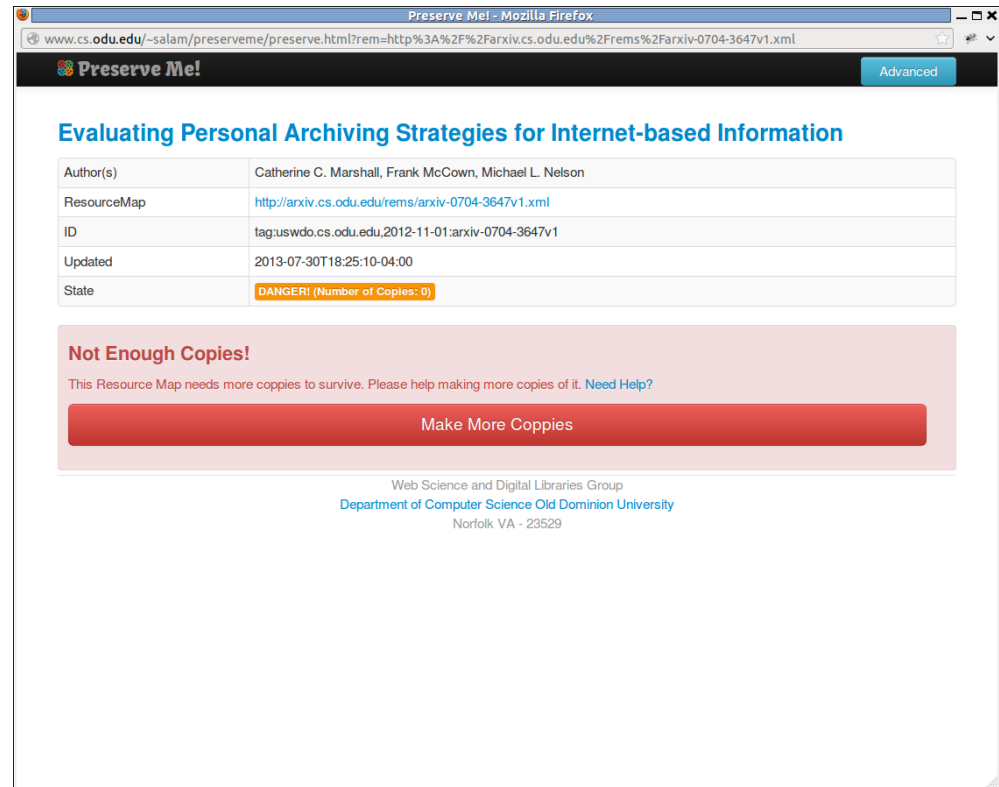
DBLP - CS Bibliography
listing | bibtex
Catherine C. Marshall
Frank McCown
Michael L. Nelson

Bookmark (what is this?)

```
...  
<link rel="resourcemap" type="application/atom+xml;type=entry"  
href="http://arxiv.cs.odu.edu/rems/arxiv-0704-3647v1.xml" />  
<link rel="aggregation" href="http://arxiv.cs.odu.edu/rems/arxiv-0704-  
3647v1.xml#aggregation" />  
<script src="http://www.cs.odu.edu/~salam/wsd/wsd/work/preserveme.js"></script>  
...
```

USW algorithm popup

- Written in JavaScript
- Relies on domain services
 - Copy -> creates copy of a WO
 - Edit -> update own REM
- Uses communications mechanism based on Sawood Alam's master's thesis



The screenshot shows a web browser window titled "Preserve Me! - Mozilla Firefox" with the URL www.cs.odu.edu/~salam/preserveme/preserve.html?rem=http%3A%2F%2Farxiv.cs.odu.edu%2Frem%2Farxiv-0704-3647v1.xml. The page header includes the "Preserve Me!" logo and an "Advanced" button. The main content area displays the title "Evaluating Personal Archiving Strategies for Internet-based Information" and a table with the following data:

Author(s)	Catherine C. Marshall, Frank McCown, Michael L. Nelson
ResourceMap	http://arxiv.cs.odu.edu/rem/arxiv-0704-3647v1.xml
ID	tag:uswdo.cs.odu.edu,2012-11-01:arxiv-0704-3647v1
Updated	2013-07-30T18:25:10-04:00
State	DANGER! (Number of Copies: 0)

Below the table, a red warning box states "Not Enough Copies!" and "This Resource Map needs more copies to survive. Please help making more copies of it. [Need Help?](#)" A large red button labeled "Make More Copies" is positioned below the warning. At the bottom of the page, the footer text reads: "Web Science and Digital Libraries Group, Department of Computer Science Old Dominion University, Norfolk VA - 23529".

USW copies: famine to feast

Name	Requirements
Famine	$h_{\text{cap}} < c_{\text{soft}} \leq c_{\text{hard}}$
Boundary Low	$h_{\text{cap}} = c_{\text{soft}} \leq c_{\text{hard}}$
Straddle	$c_{\text{soft}} \leq h_{\text{cap}} \leq c_{\text{hard}}$
Boundary High	$c_{\text{soft}} \leq c_{\text{hard}} = h_{\text{cap}}$
Feast	$c_{\text{soft}} \leq c_{\text{hard}} < h_{\text{cap}}$

Final states for copying policies and named conditions

