9-9-2014

When Should I Make Preservation Copies of Myself?

Charles L. Cartledge
Old Dominion University, ccartled@odu.edu

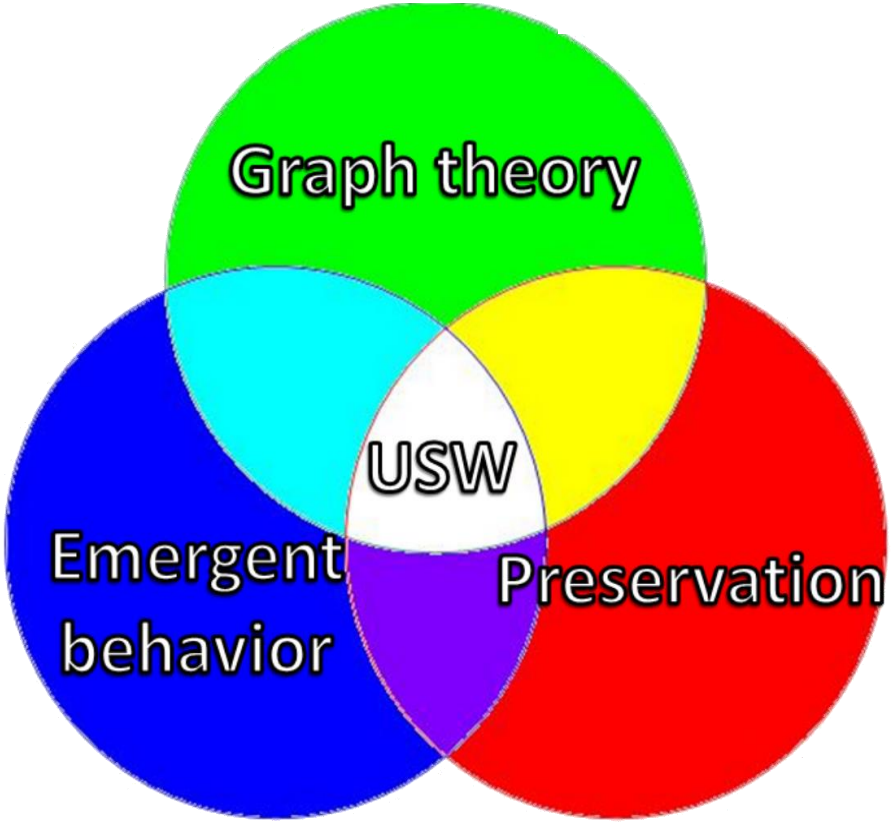
Michael L. Nelson
Old Dominion University, mnelson@odu.edu

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



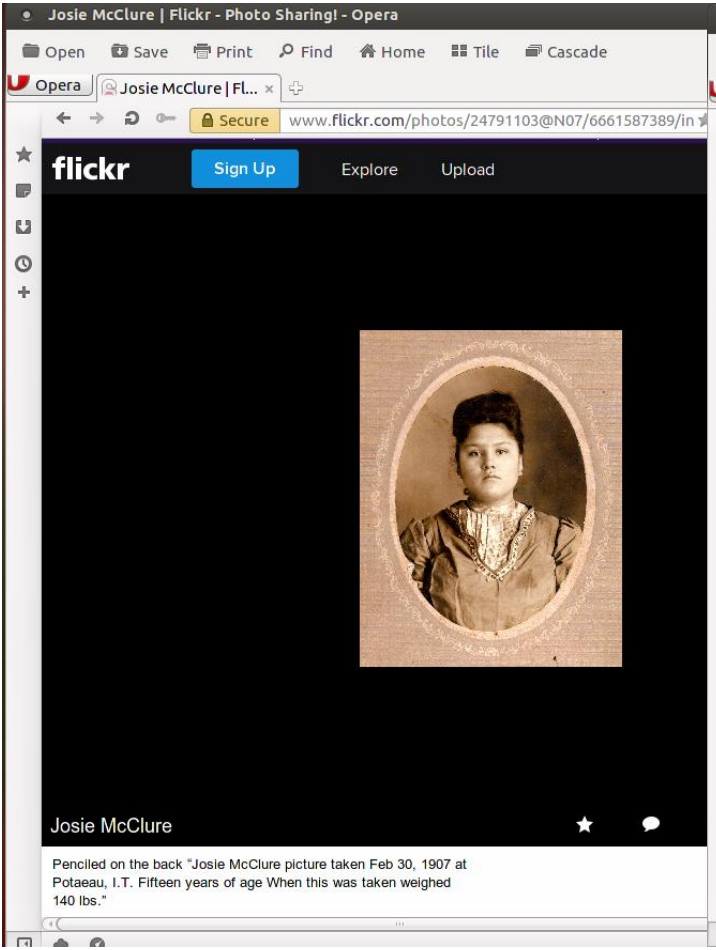
Preservation via benign neglect

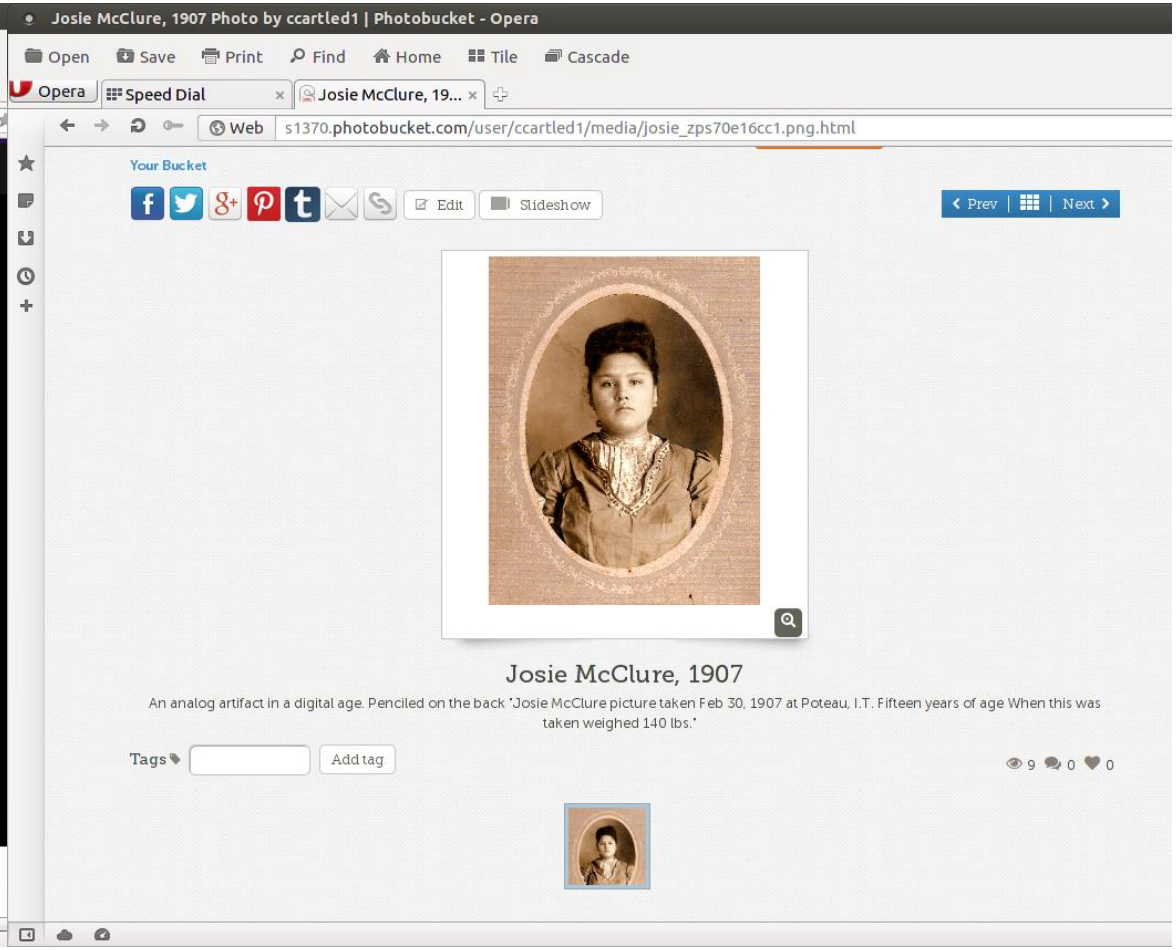
Handwritten on the back of the photo:

"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.?





Crowd sourcing preservation

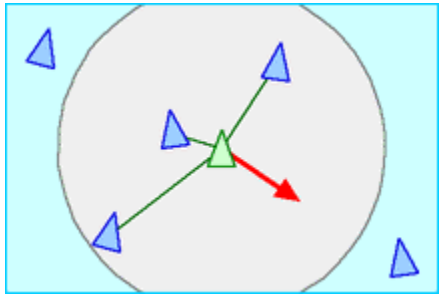
- "Everyone is a curator …"
  - Crowd sourced activity
  - Unscheduled
  - Willing to wait a long time
- Enlist humans in creation and maintenance – opposite of benign neglect

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

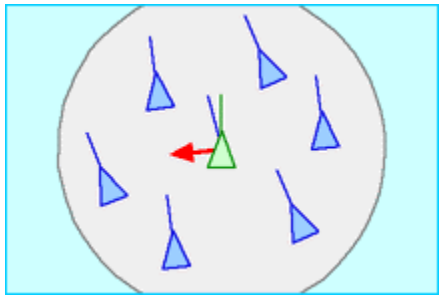
See also: http://ws-dl.blogspot.com/2013/10/2013-10-23-preserve-me-if-you-can-using.html

Emergent behavior: flocking boids

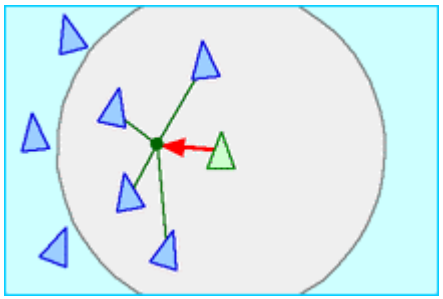
- Craig Reynolds – basis of herd and flock behavior in computer animations
  - 3 rules
    - Collision avoidance
    - Velocity matching
    - Flock centering
  - No central control, everything based on local knowledge only
- **Simple rules produce complex, emergent behavior**



Collision avoidance



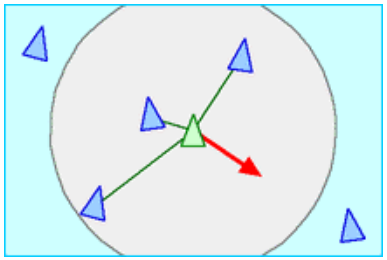
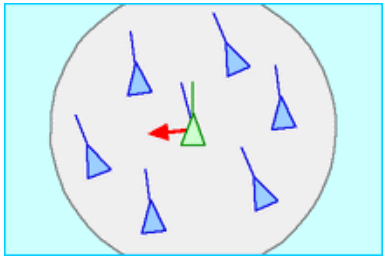
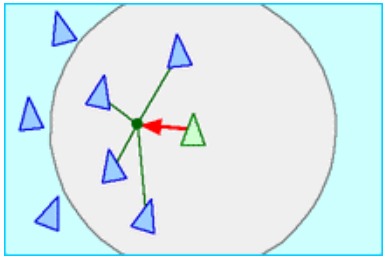
Velocity matching



Flock centering

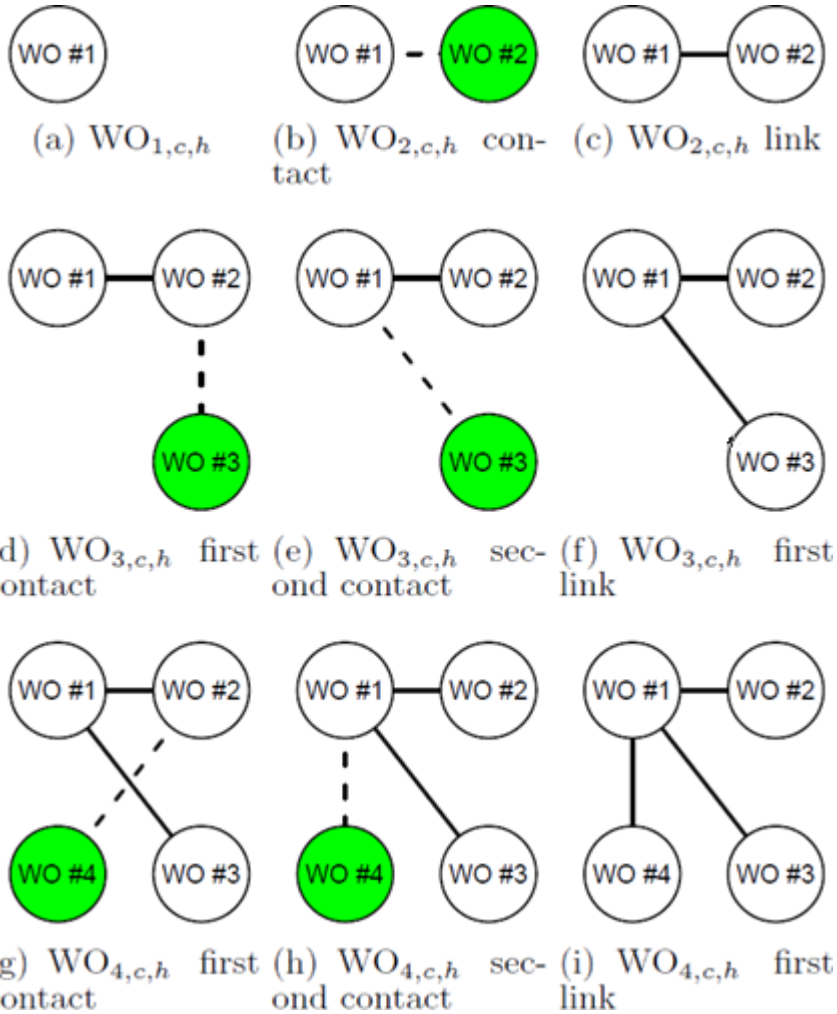
Craig W. Reynolds, Computer Animation with Scripts and Actors, ACM SIGGRAPH, vol. 16, ACM, 1982, pp. 289 - 296.
Images http://www.red3d.com/cwr/boids/

USW interpretation of flocking

Craig Reynolds' "boids"		USW interpretation
Collision avoidance		**Each WO has a unique URI**
Velocity matching		**Matching number of copies/family members**
Flock centering		**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**



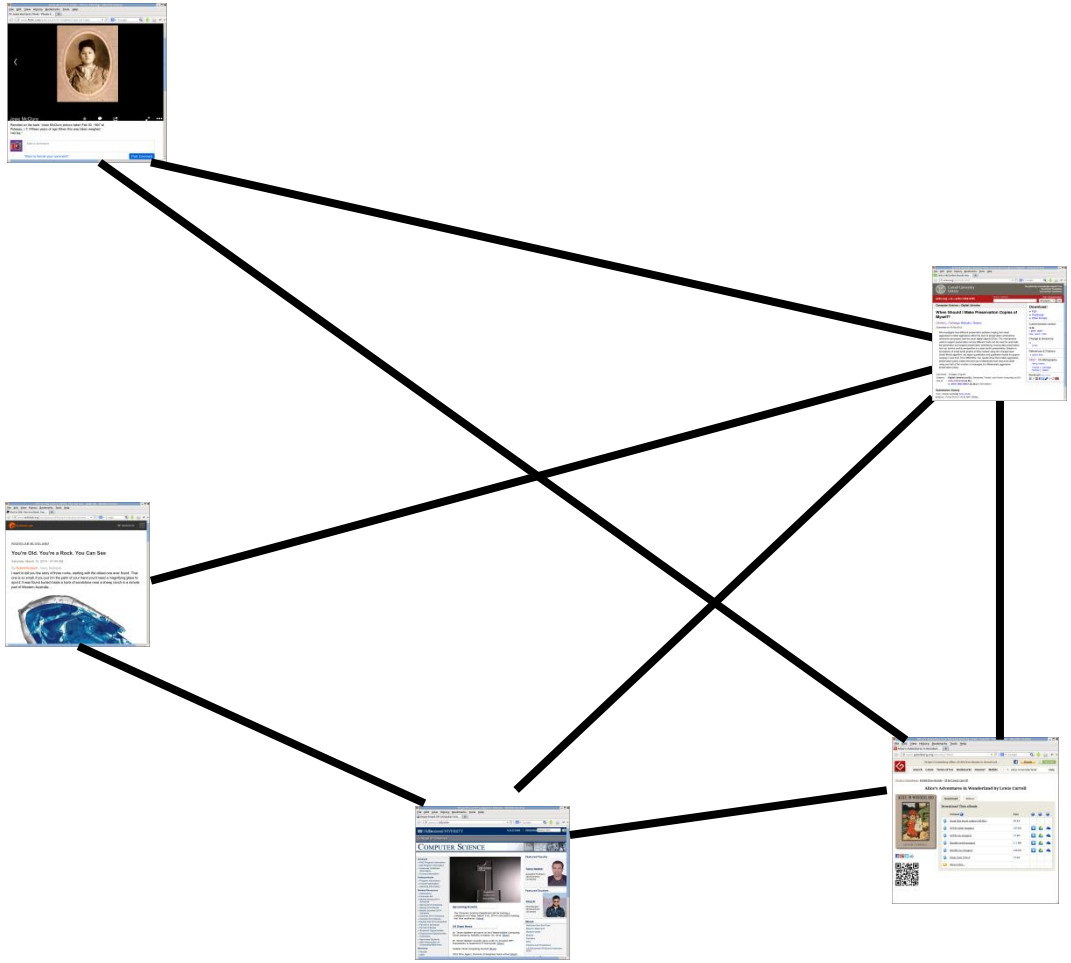
(a) $WO_{1,c,h}$ (b) $WO_{2,c,h}$ contact (c) $WO_{2,c,h}$ link

(d) $WO_{3,c,h}$ first contact (e) $WO_{3,c,h}$ second contact (f) $WO_{3,c,h}$ first link

(g) $WO_{4,c,h}$ first contact (h) $WO_{4,c,h}$ second contact (i) $WO_{4,c,h}$ first link

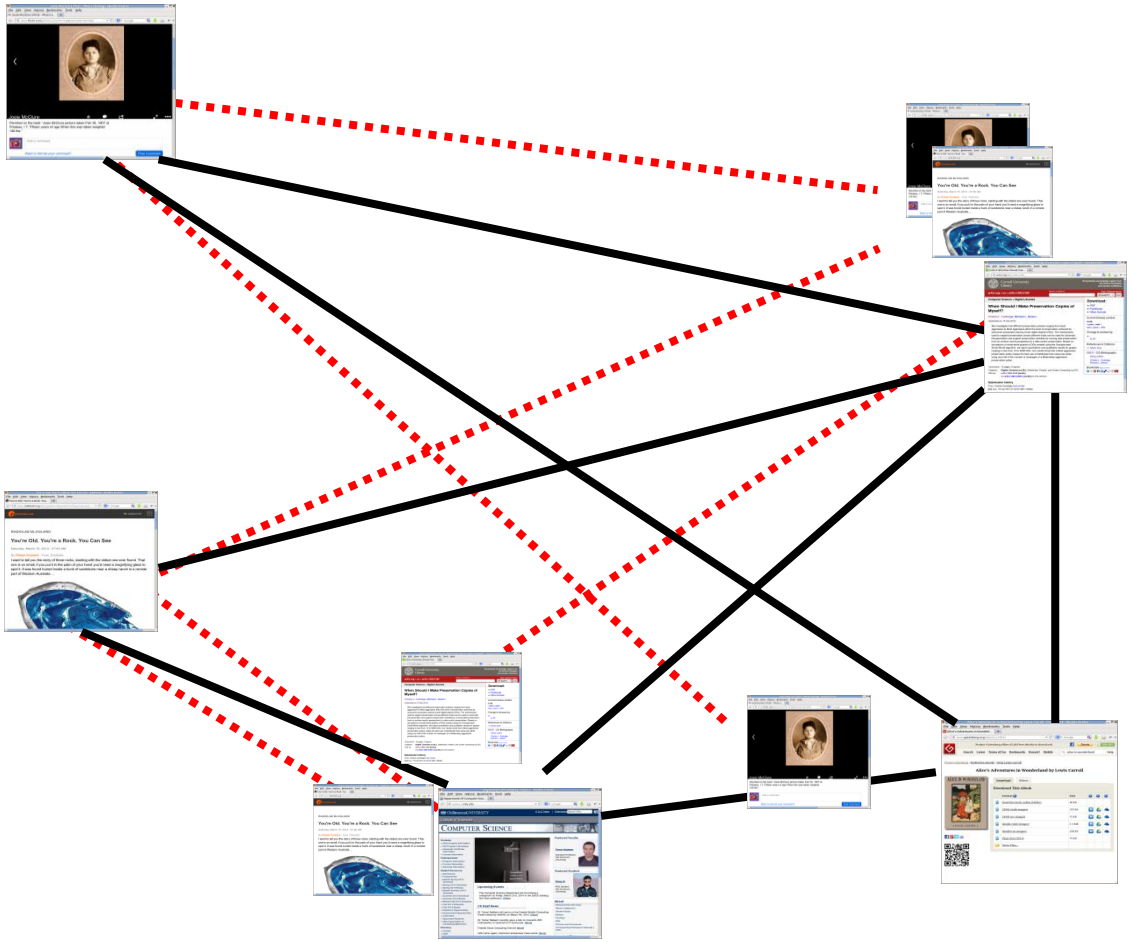
USW WO “friendship” links

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



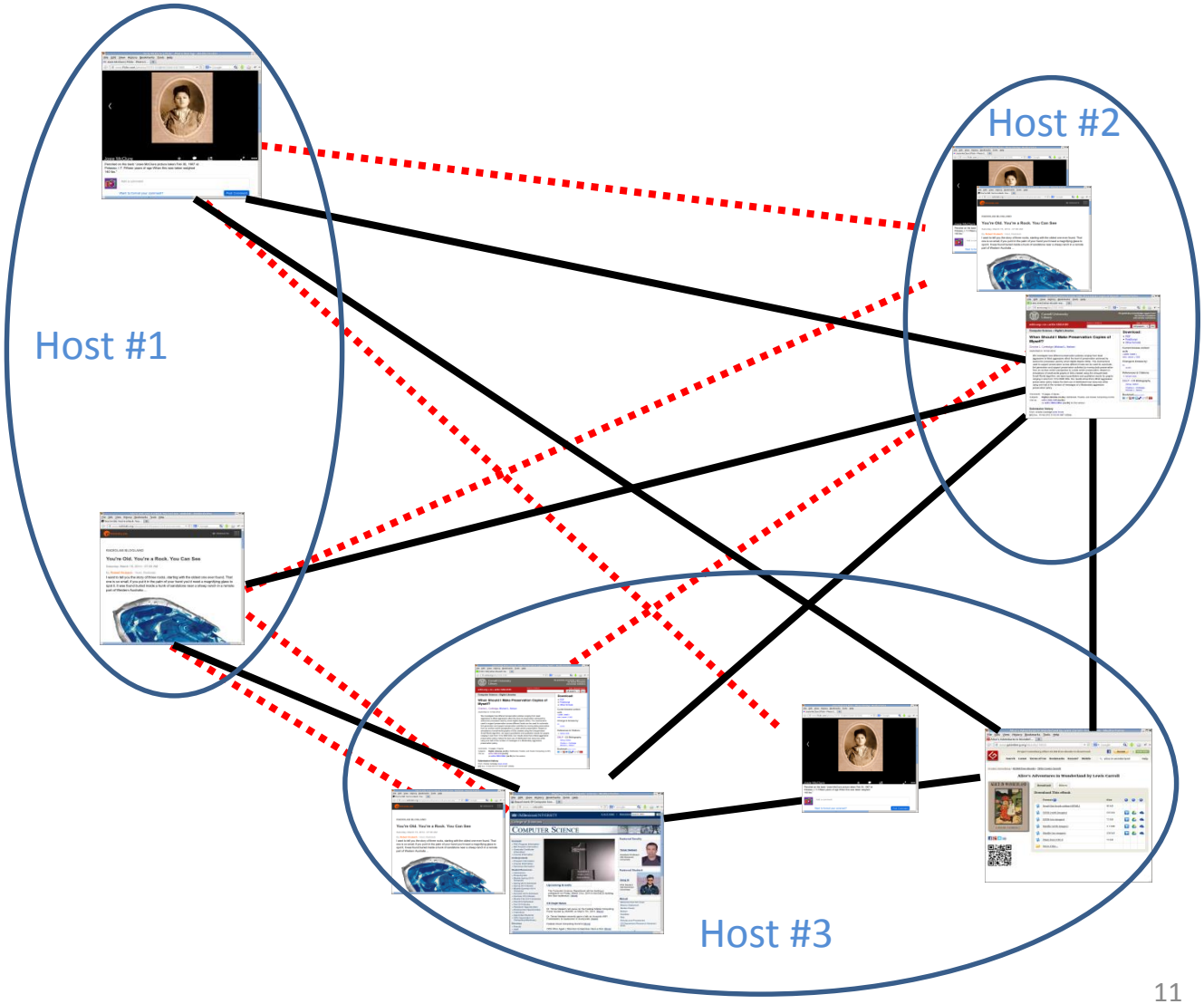
USW WO “families”

A family is a set of copies of the same WO



USW hosts

Family members live on different hosts



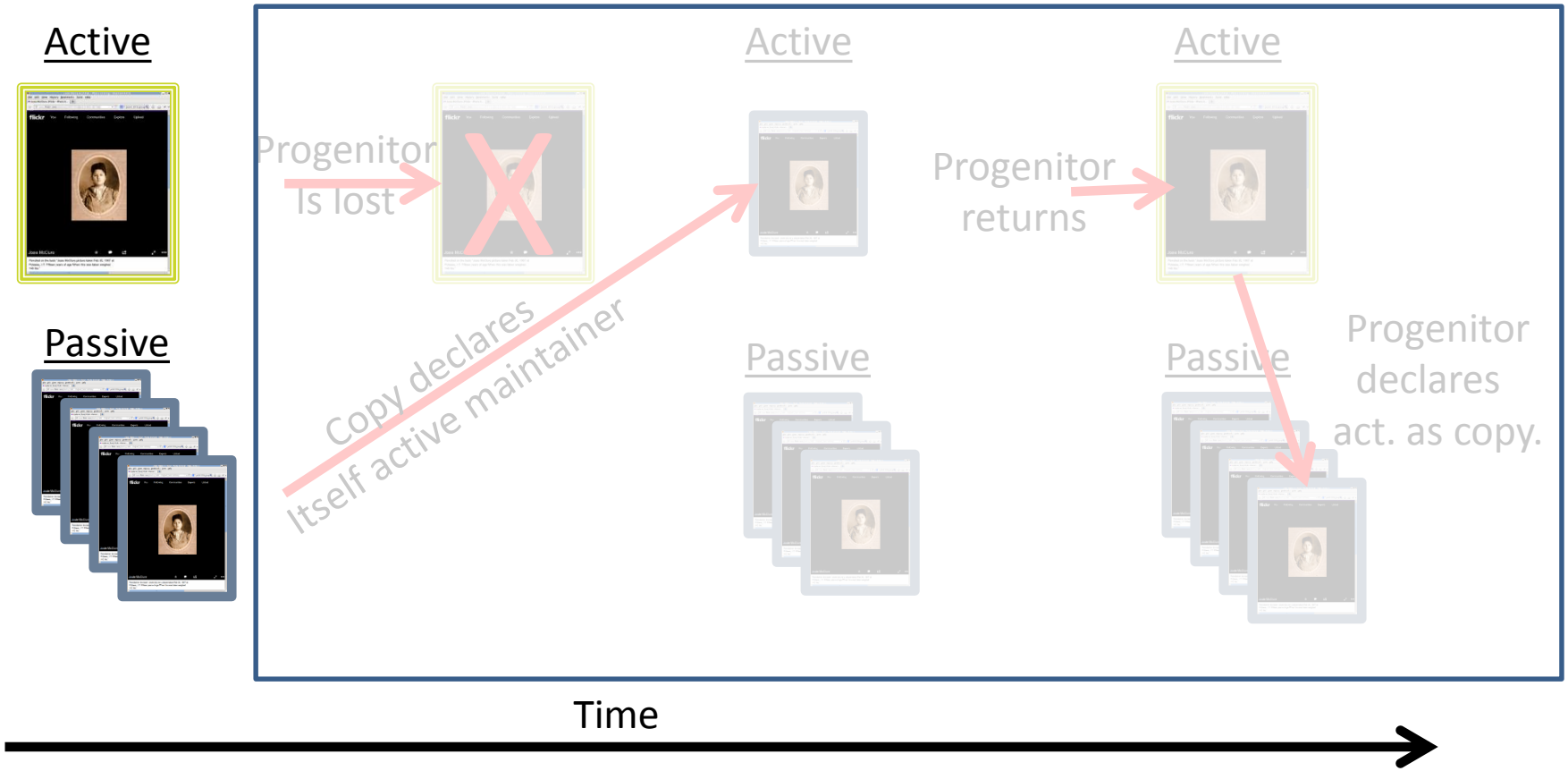
WO roles & responsibilities

- 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

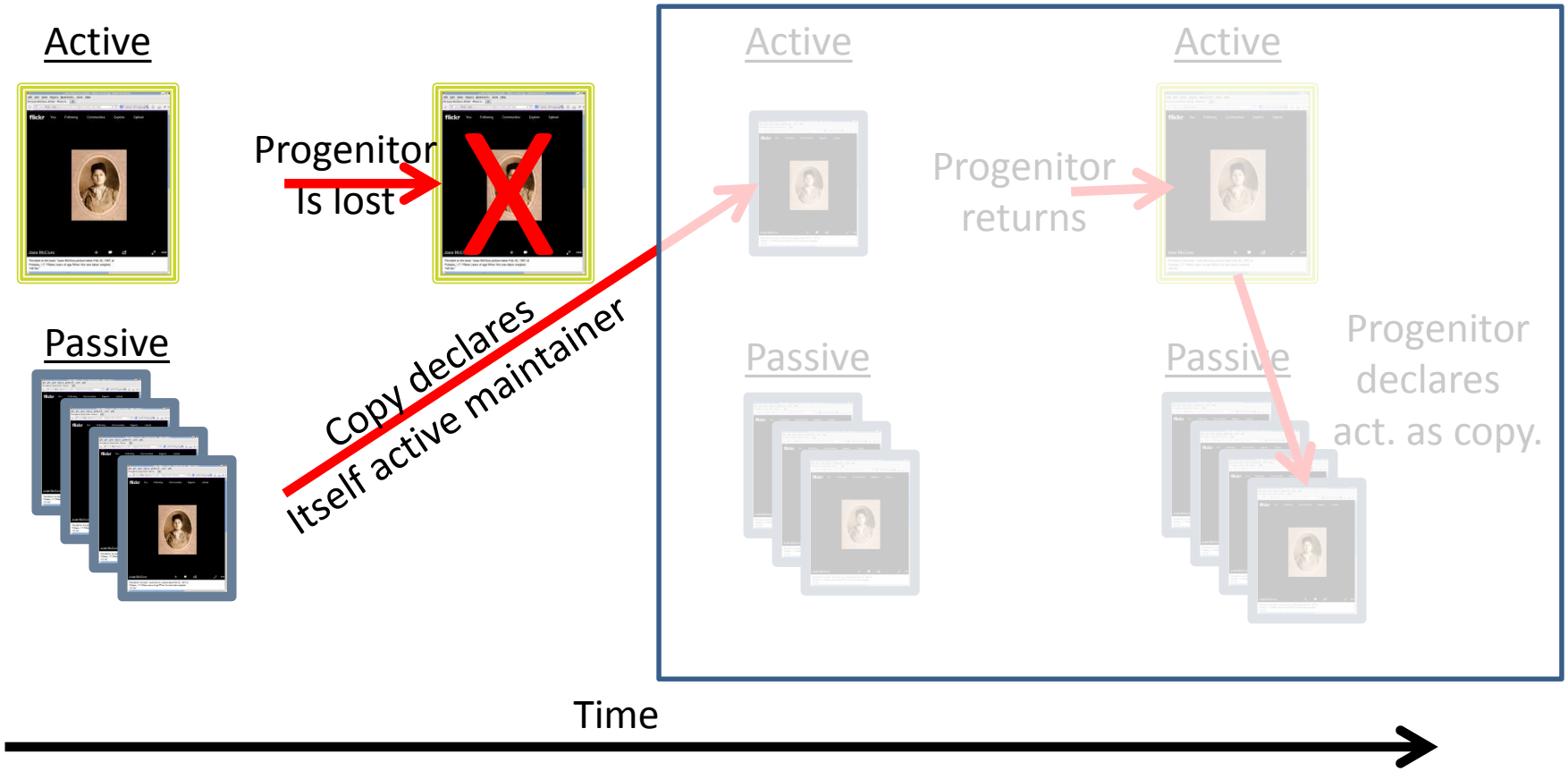

Copies

Active and passive maintenance activities



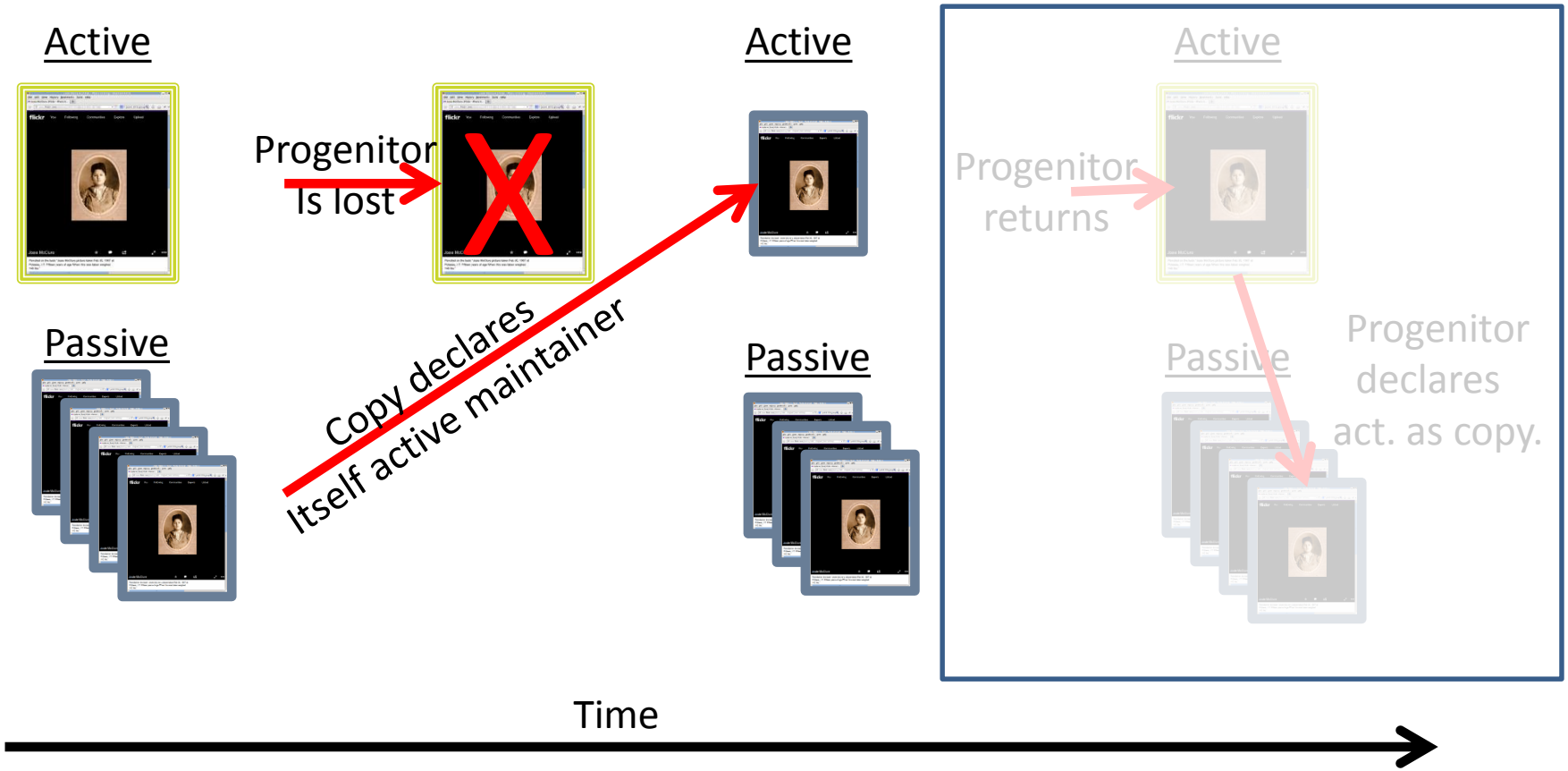
- Active maintainer (the WO with earliest timestamp) – currently charged with making copies and related housekeeping
- Passive maintainer – all other WOs

Progenitor is lost



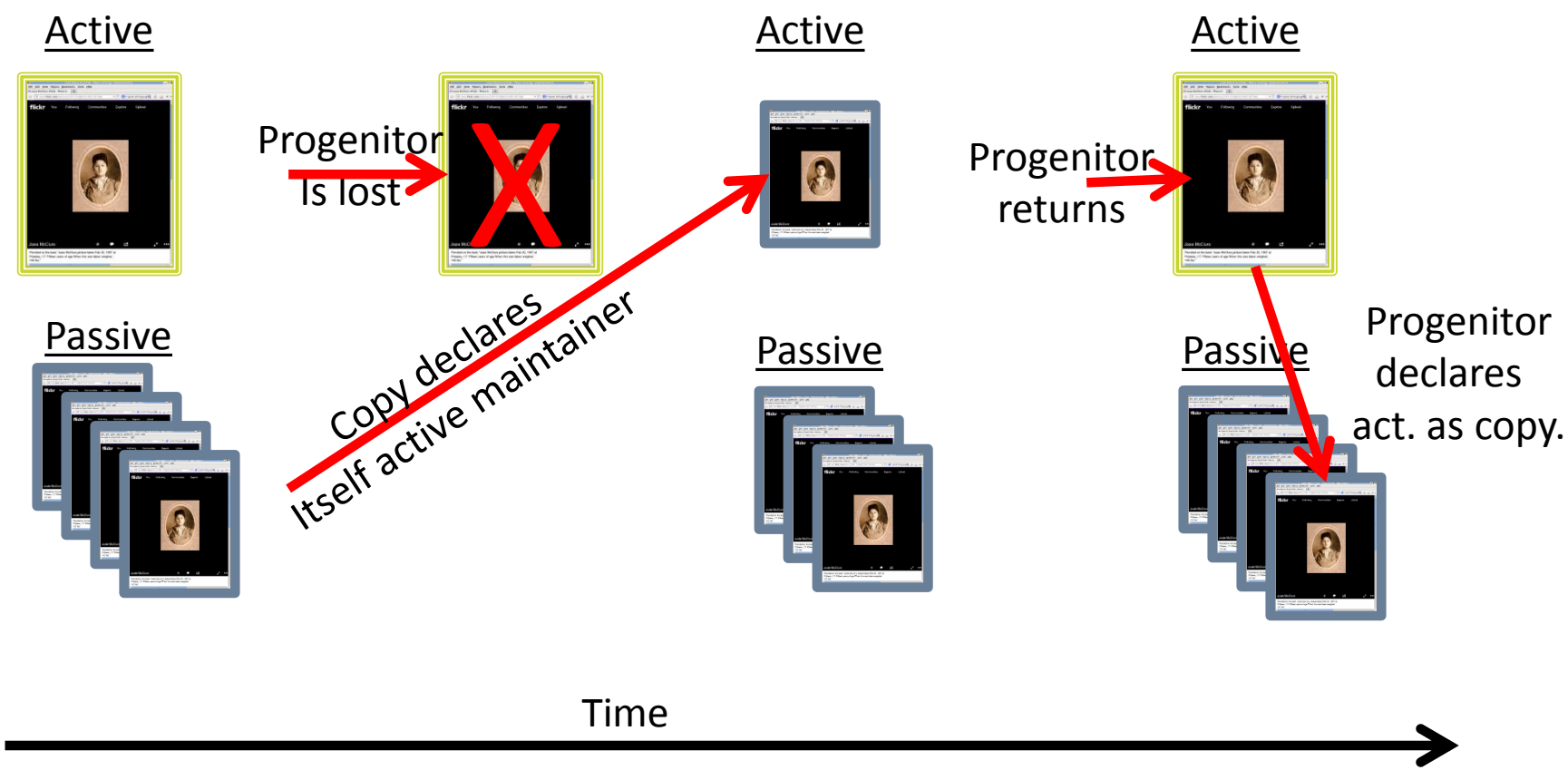
- Active maintainer – currently charged with making copies and related housekeeping
- Passive maintainer – all other WOs

A new active maintainer



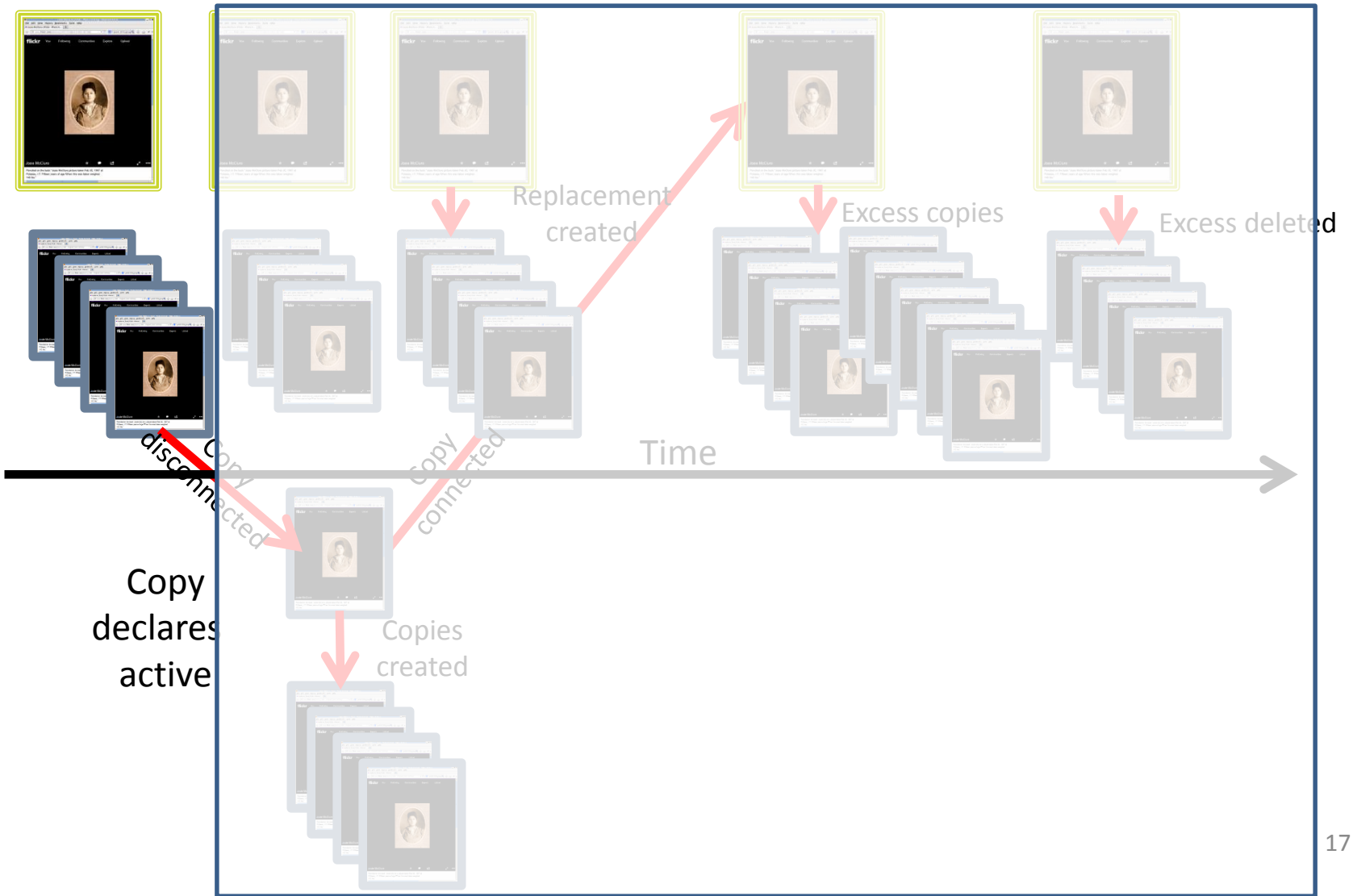
- Active maintainer – currently charged with making copies and related housekeeping
- Passive maintainer – all other WOs

Progenitor returns and assumes active maintainer role

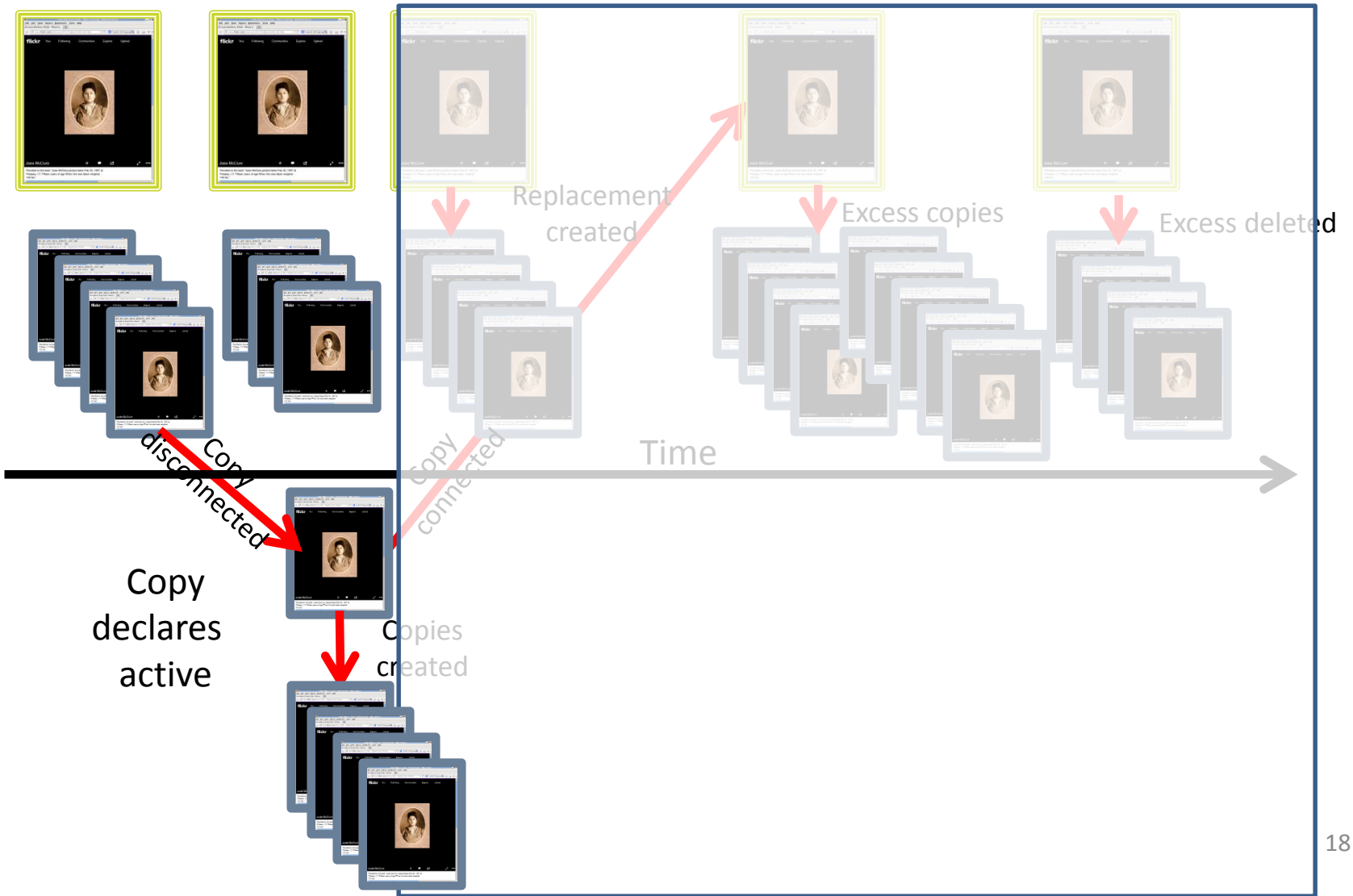


- Active maintainer – currently charged with making copies and related housekeeping
- Passive maintainer – all other WOs

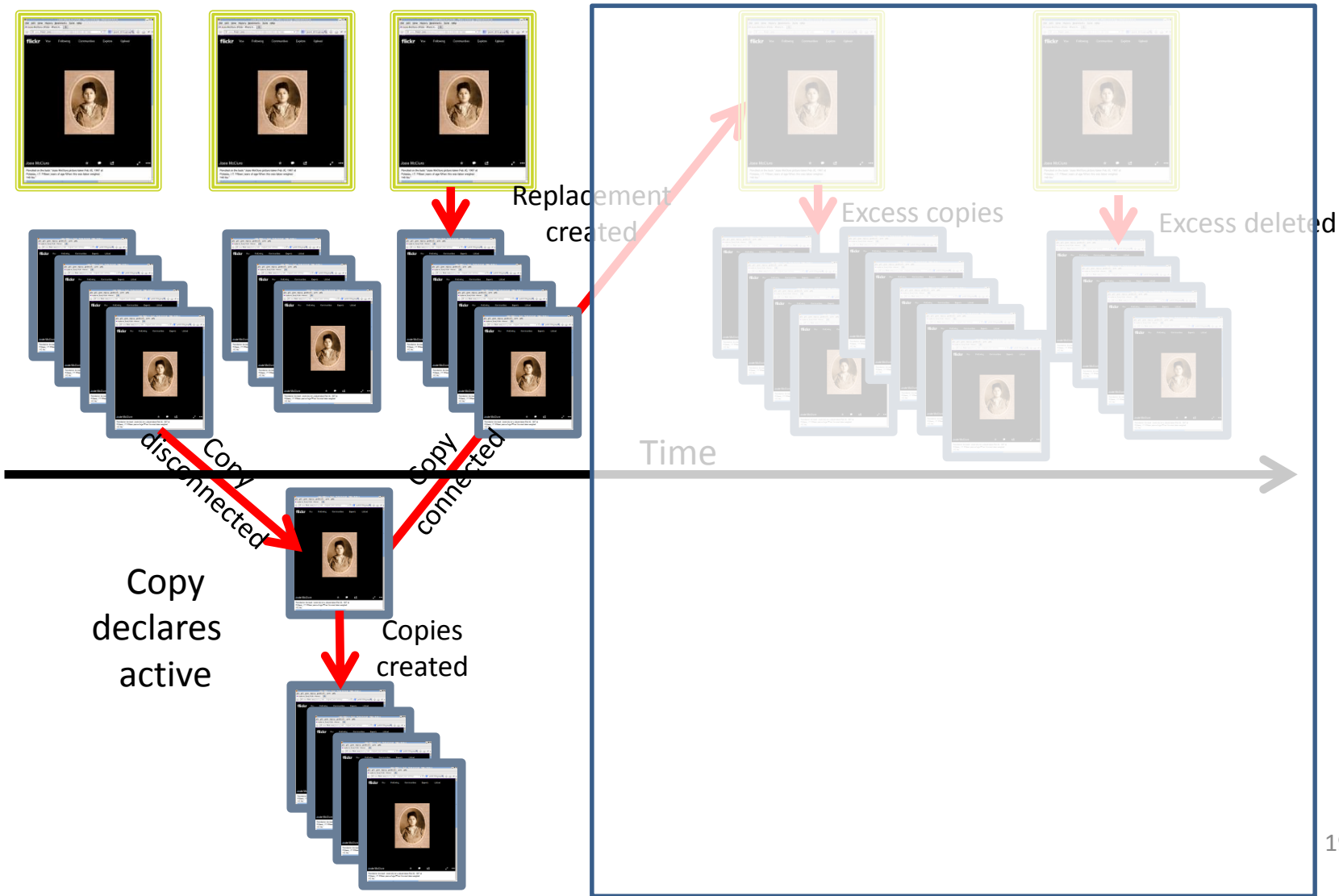
Progenitor has made copies



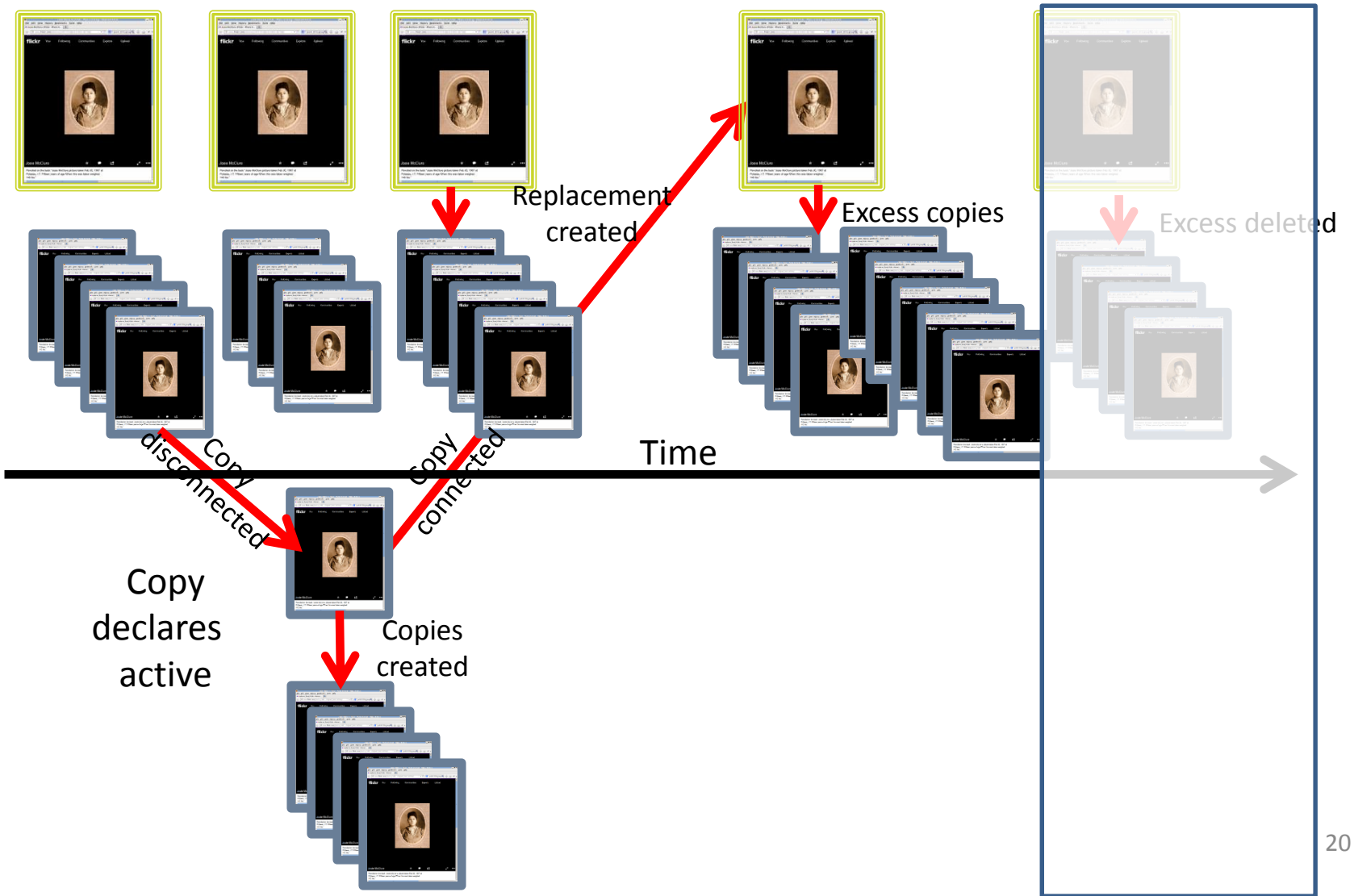
A copy is disconnected from the family



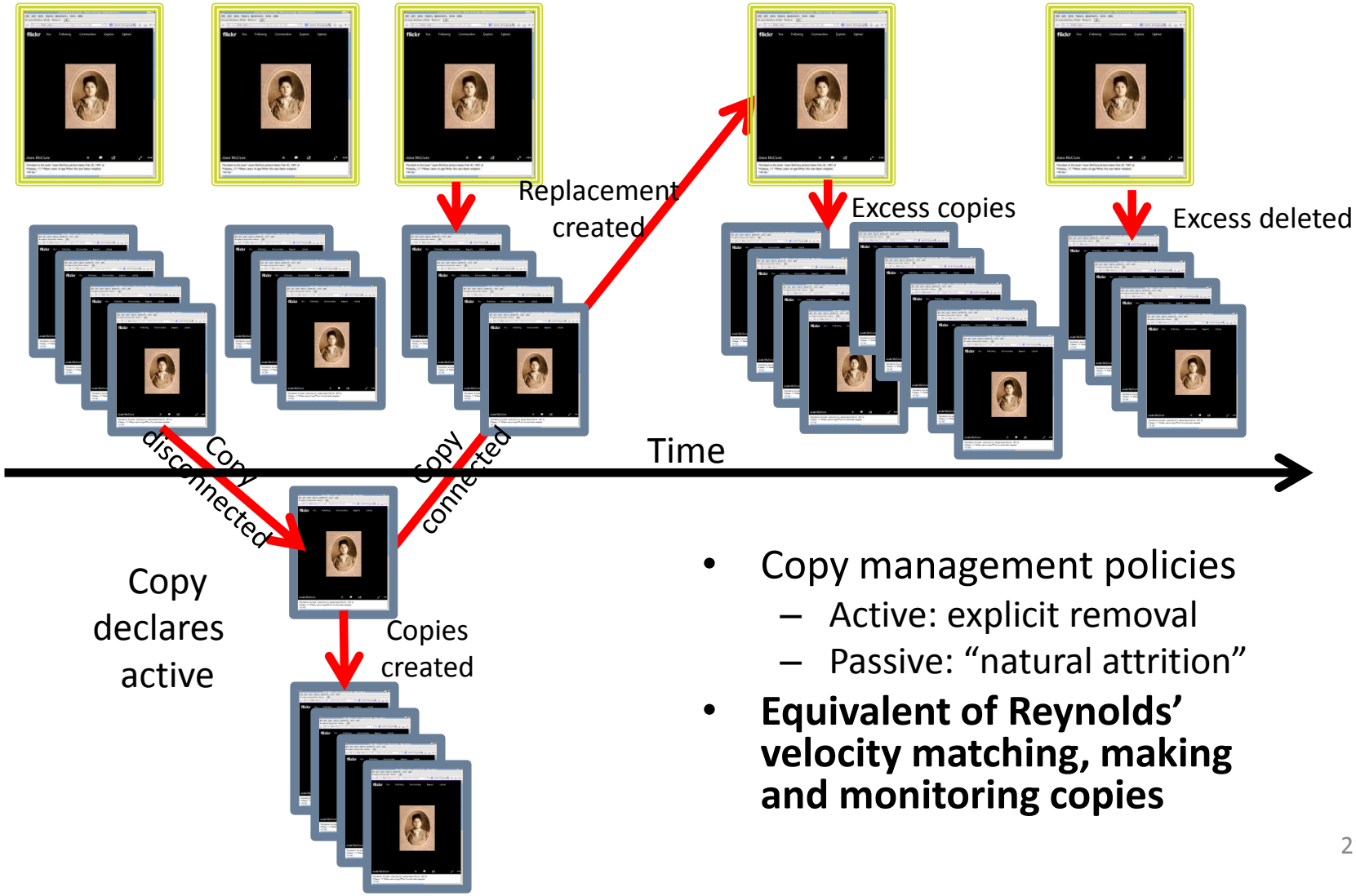
Two active maintainers make copies



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**

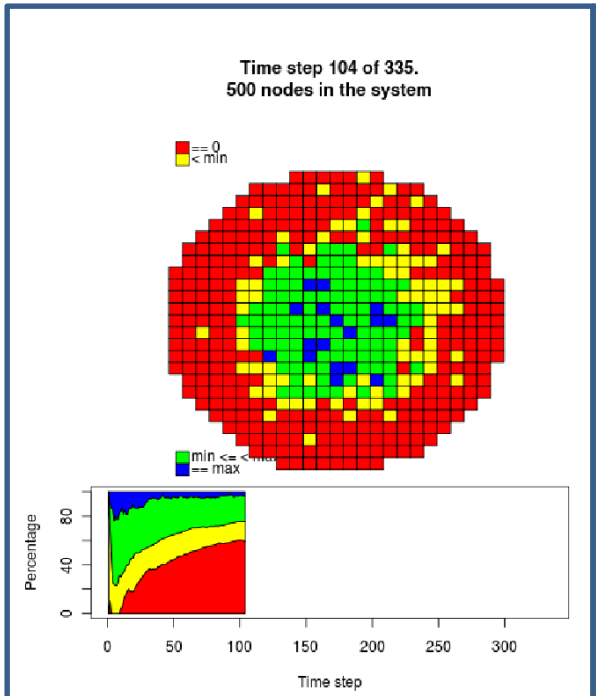
Parameters

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

Three USW copying policies

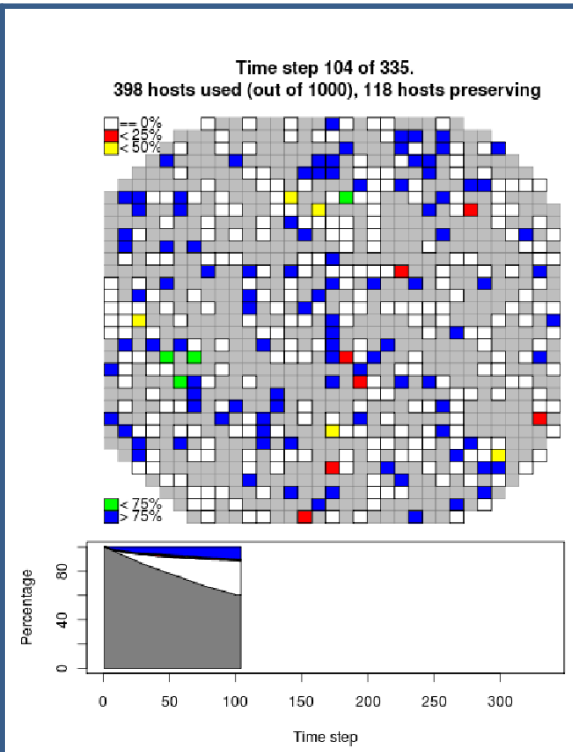
- **Least aggressive** – one at a time to c_{hard}
- **Moderately aggressive** – as quickly as possible to c_{soft} and then one at a time c_{hard}
- **Most aggressive** – as quickly as possible to c_{hard}
- Constraints:
  - WOs can only take action when woken up by interactive users or other WOs (i.e., 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



WOs preservation status

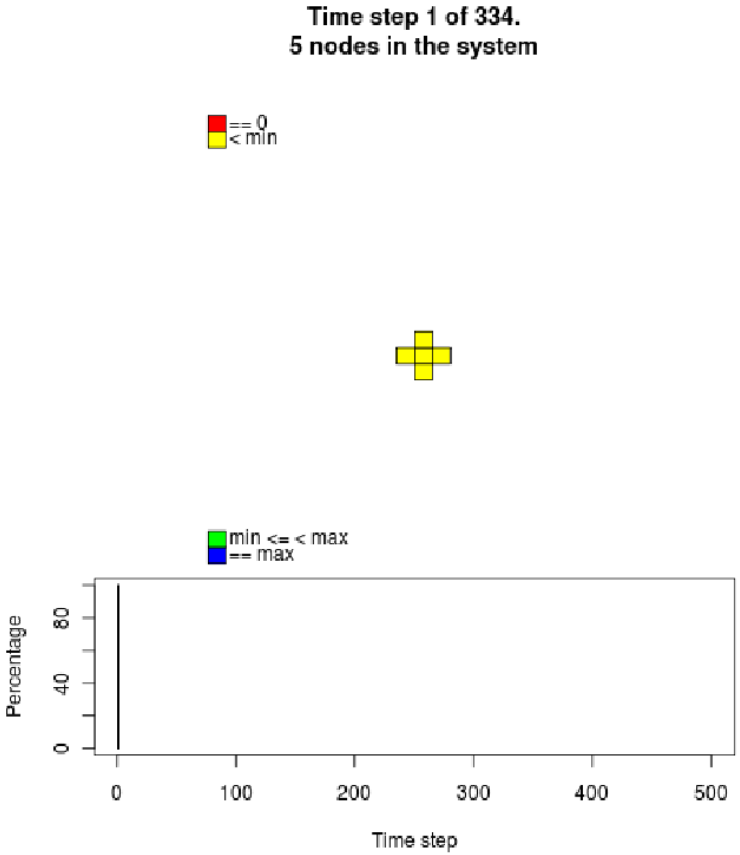
- None
- $< C_{soft}$
- $C_{soft} <= N < C_{hard}$
- $N == C_{hard}$

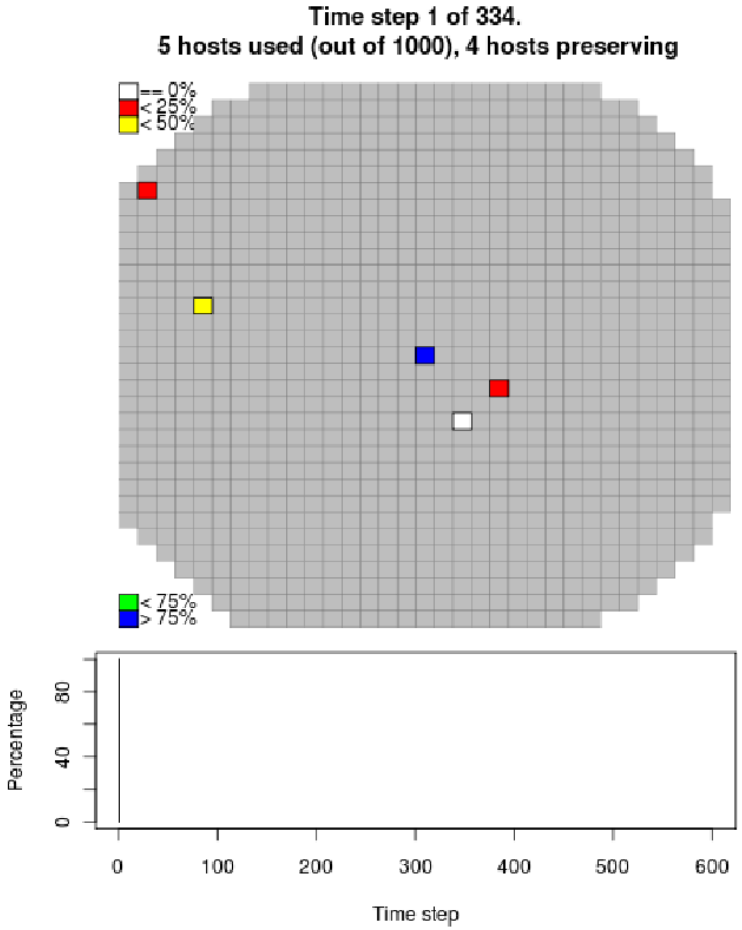


Hosts utilization status

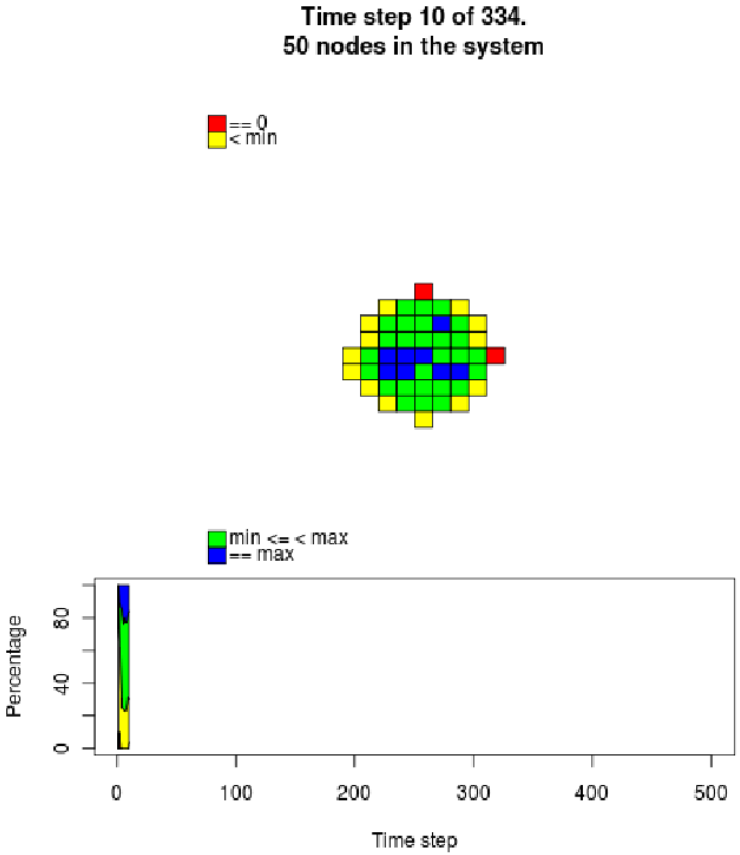
- 0%
- < 25%
- < 50 %
- < 75%
- > 75%

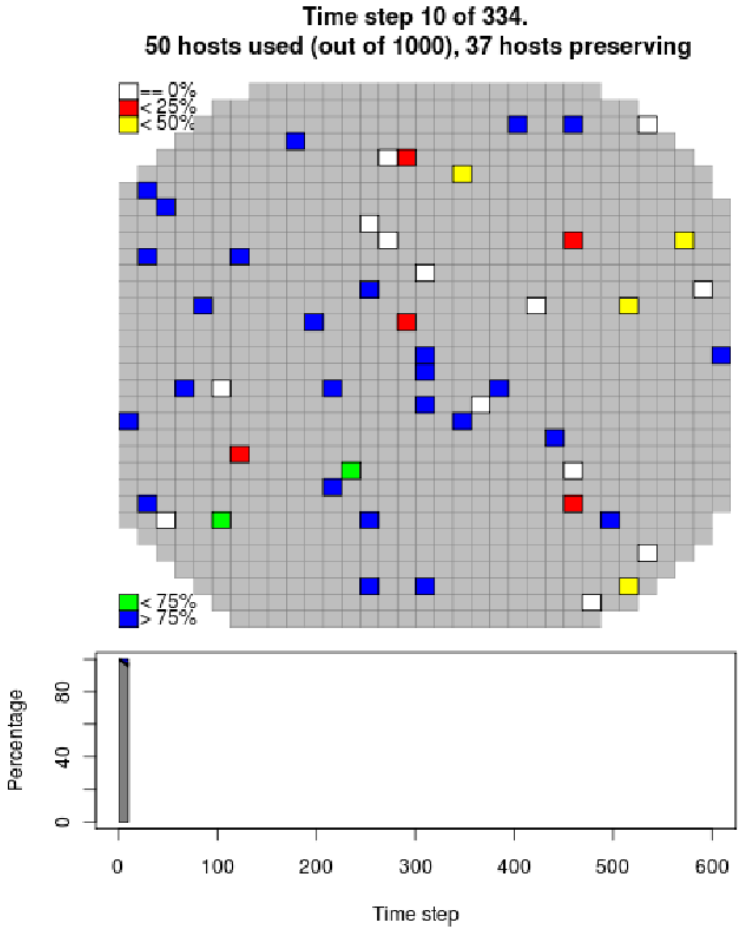
Least aggressive (t = 1)



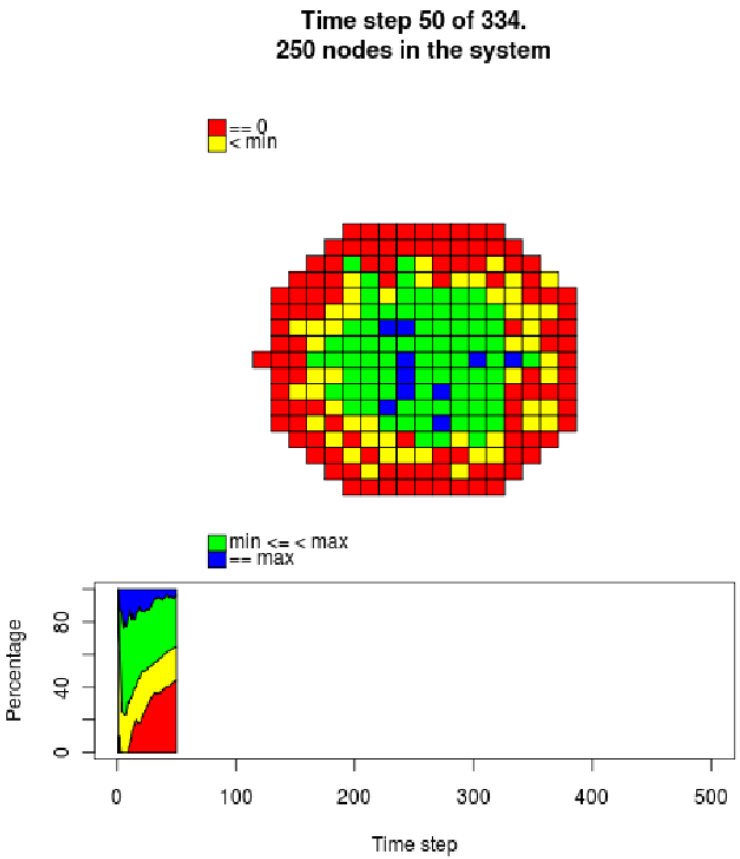


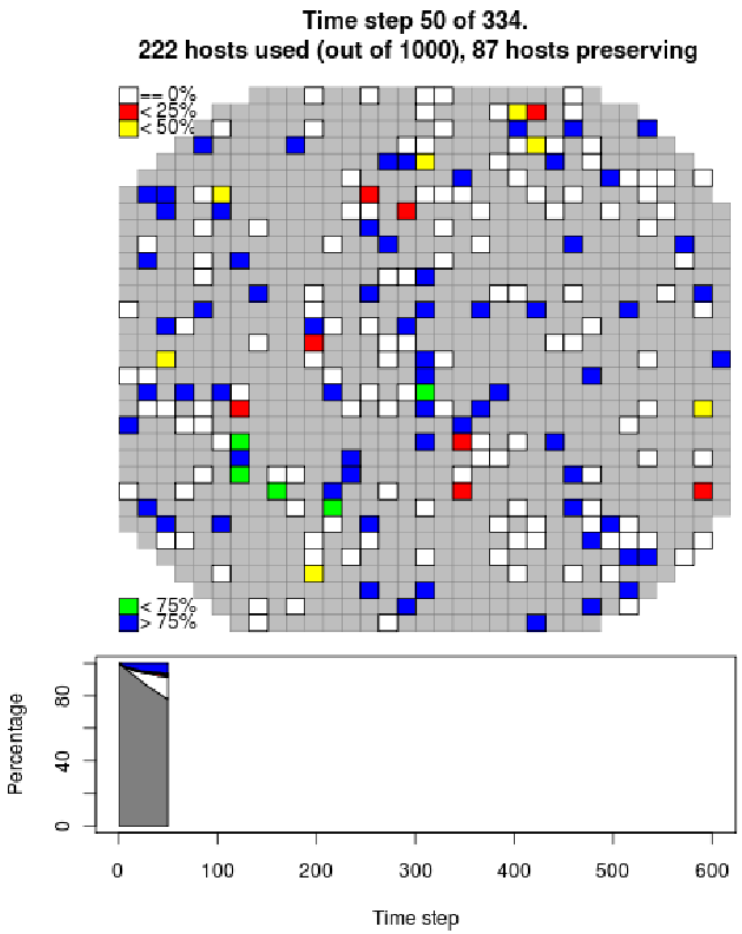
Least aggressive (t = 10)



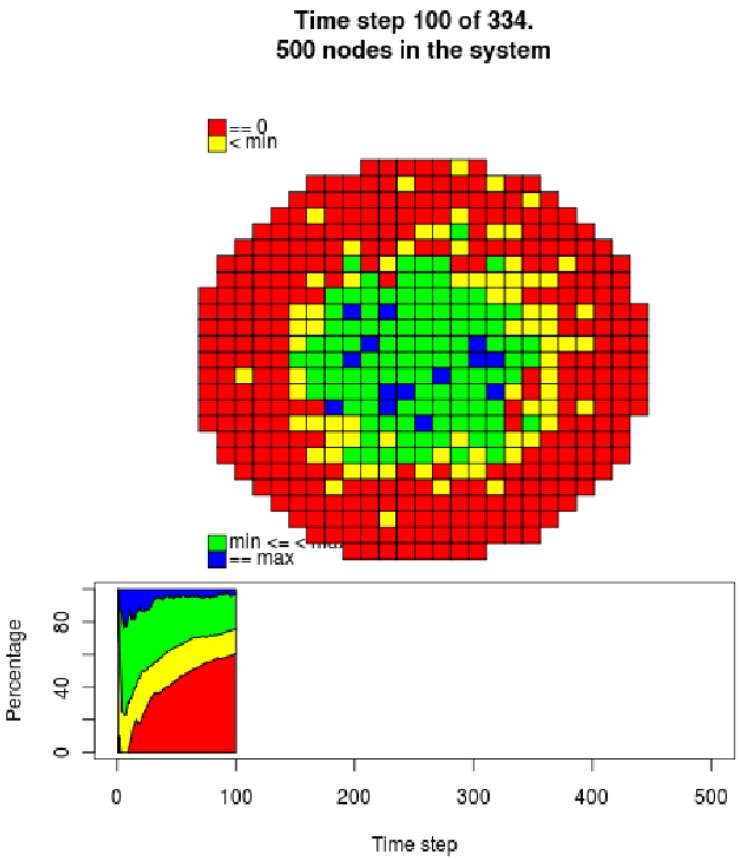


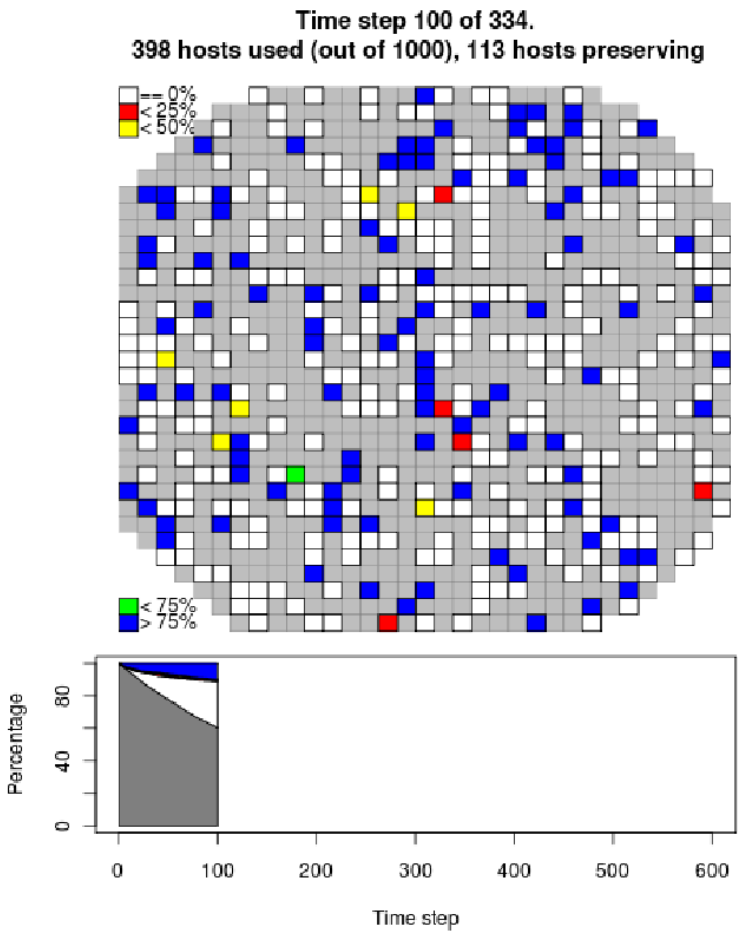
Least aggressive (t = 50)





Least aggressive (t = 100)

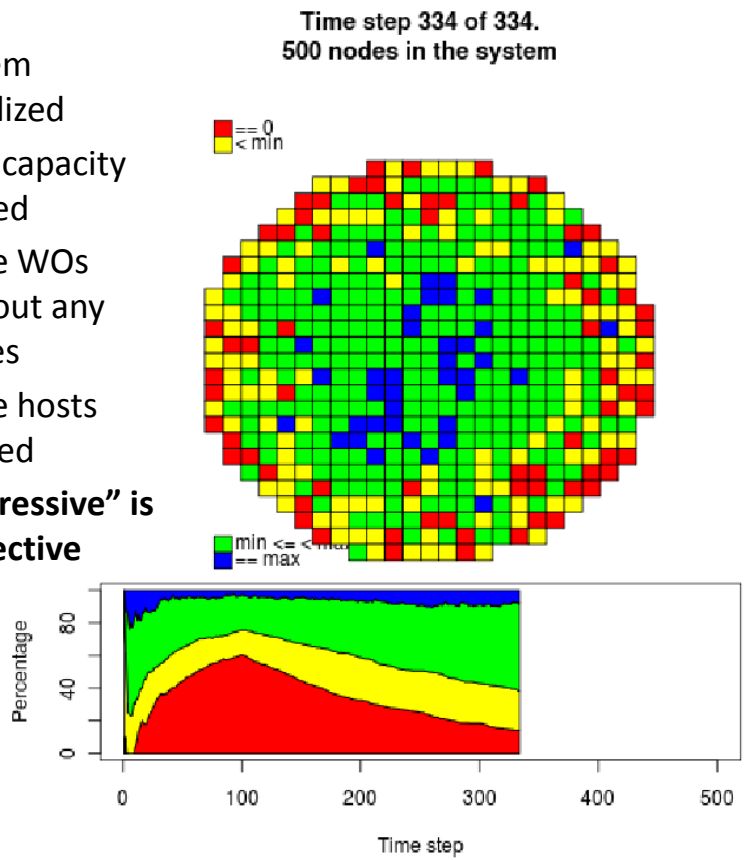


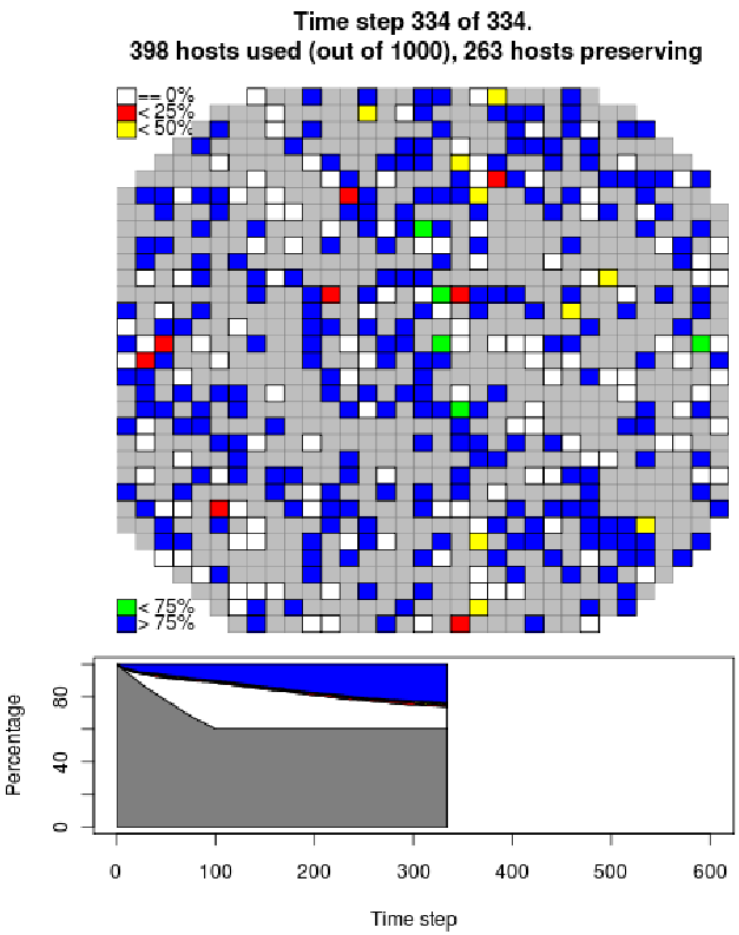


- A full YouTube video is available at: http://youtu.be/sHJGYphqtK4

Least aggressive (final)

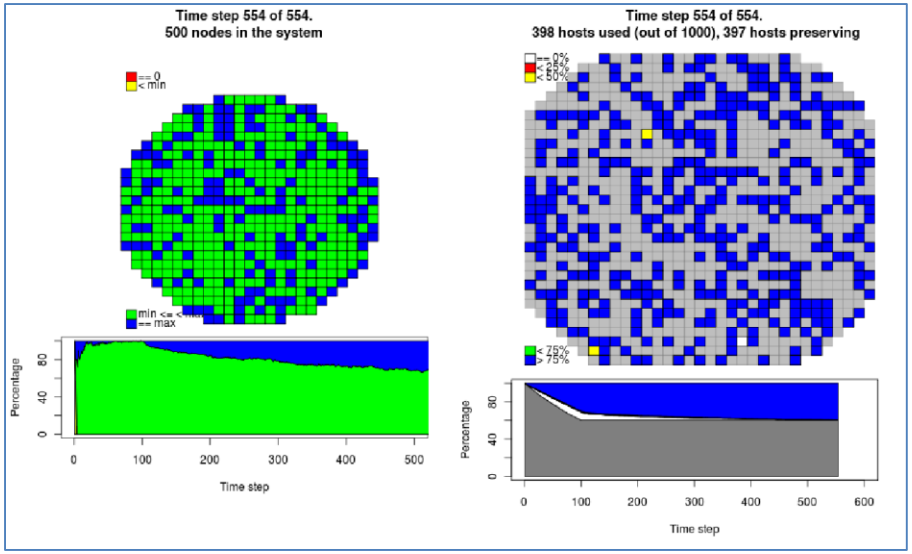
- Results
  - System stabilized
  - Host capacity limited
  - Some WOs without any copies
  - Some hosts unused
- **"Least aggressive" is not an effective policy**

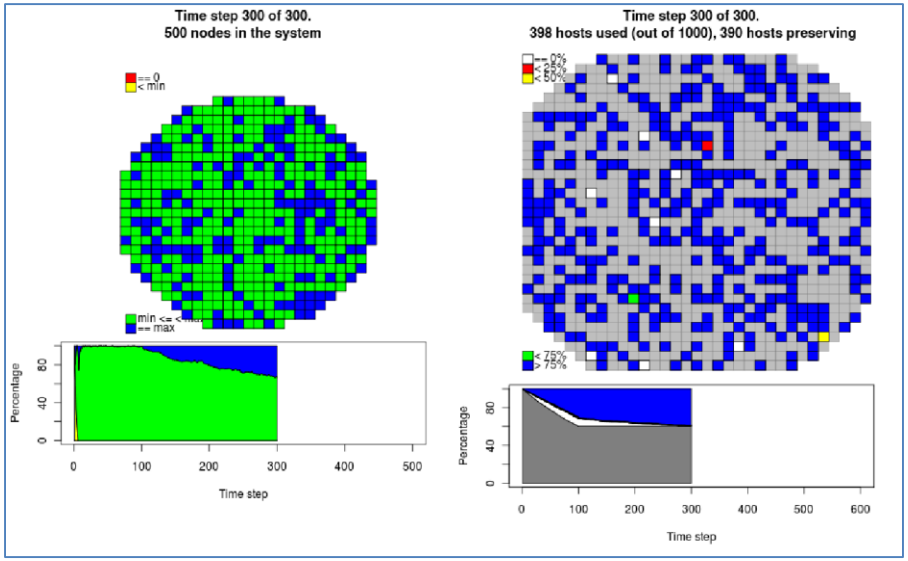




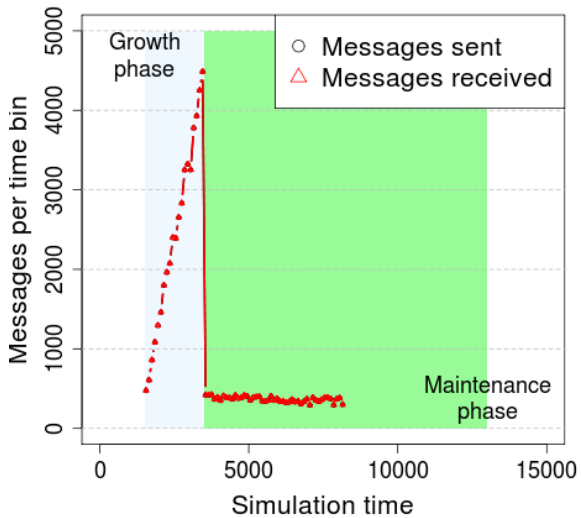
Which policy to choose?

- **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

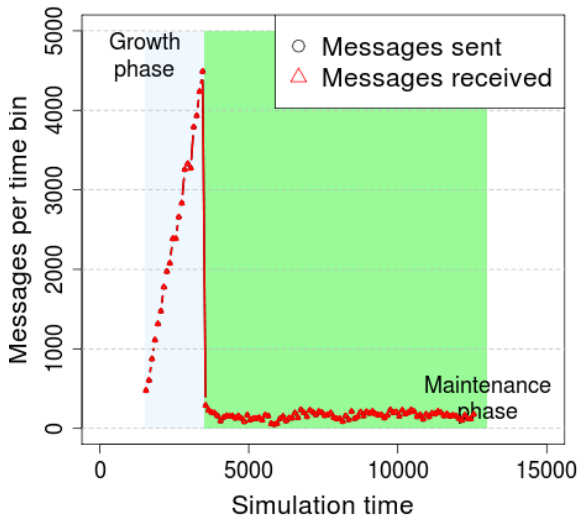




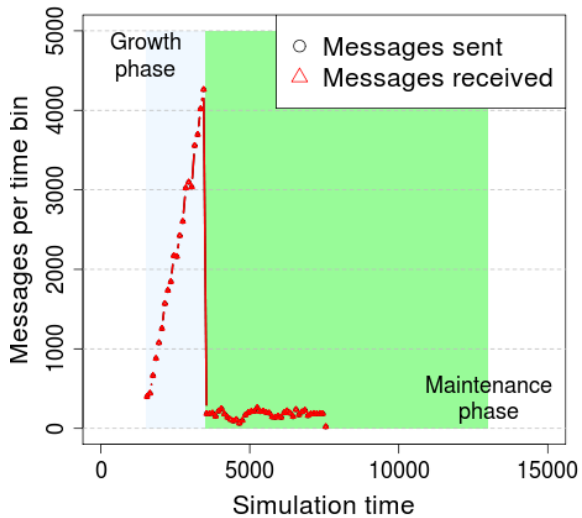
How does policy affect message exchange?



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

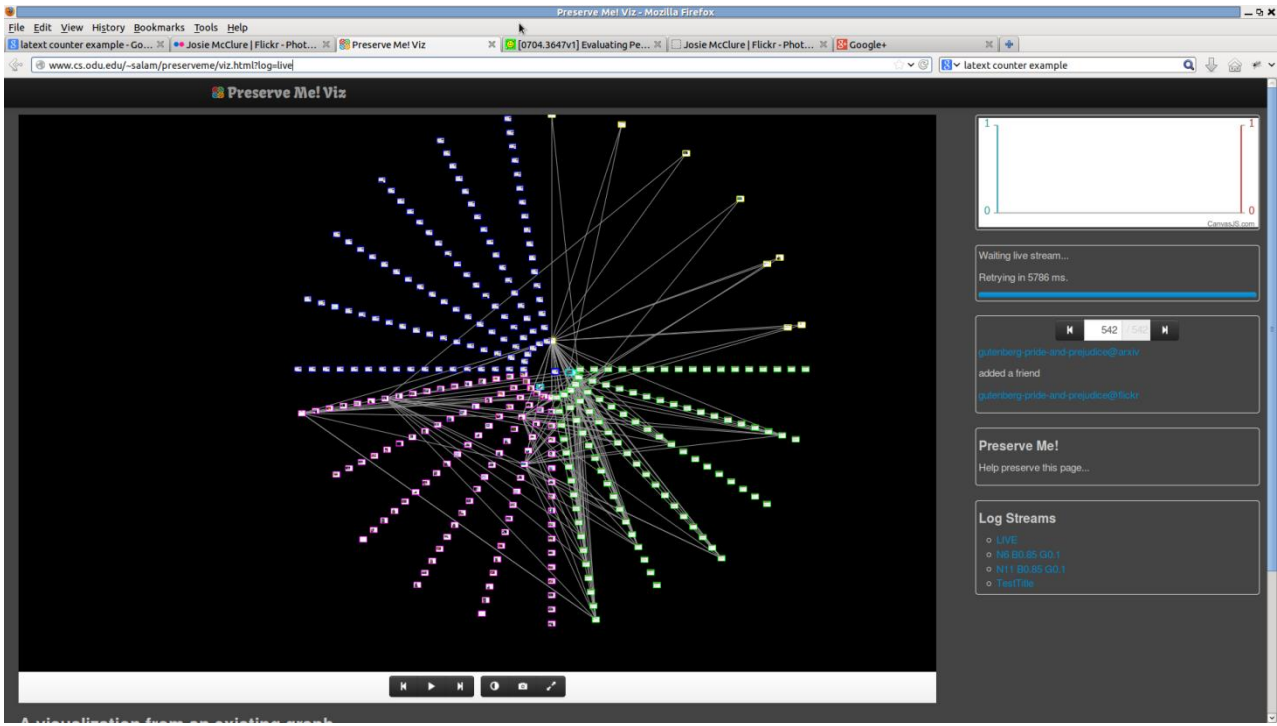
WO memory: simulated via “edit” service

```
<link rel="alternate" type="text/html" href="http://flickr.cs.odu.edu/flickr-24791
<link rel="self" type="application/atom+xml" href="http://flickr.cs.odu.edu/rems/f
<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" hr
<link rel="http://wsdl.cs.odu.edu/uswdo/terms/synchronize" type="message/http" hre
<link rel="http://wsdl.cs.odu.edu/uswdo/terms/httpmailbox#self" href="http://ws-dl
<link rel="http://wsdl.cs.odu.edu/uswdo/terms/httpmailbox#all" href="http://ws-dl-
<link rel="http://wsdl.cs.odu.edu/uswdo/terms/httpmailbox#family" href="http://ws-
<link rel="http://www.openarchives.org/ore/terms/describes" href="http://flickr.cs
```


- 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).

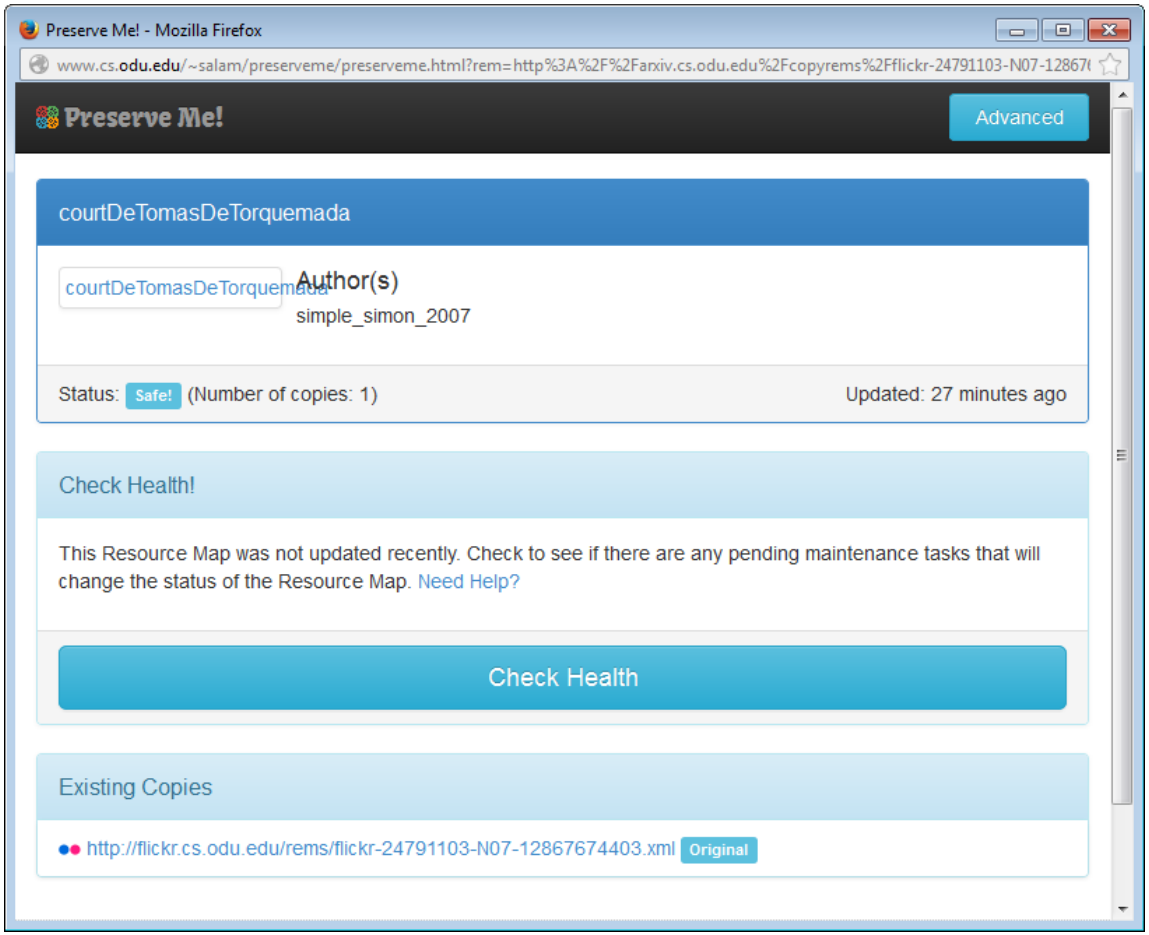
Preserve Me Viz! with new connections

- New friend connections
- New copy locations

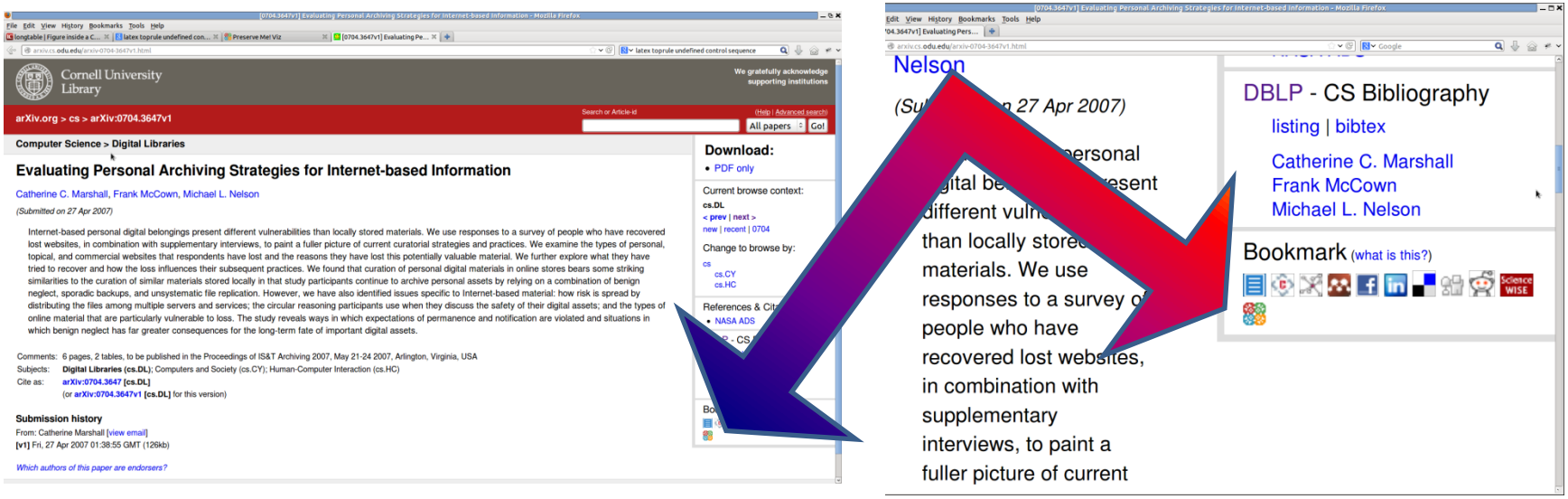


Preserve Me “Basic” on a copy

- 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.



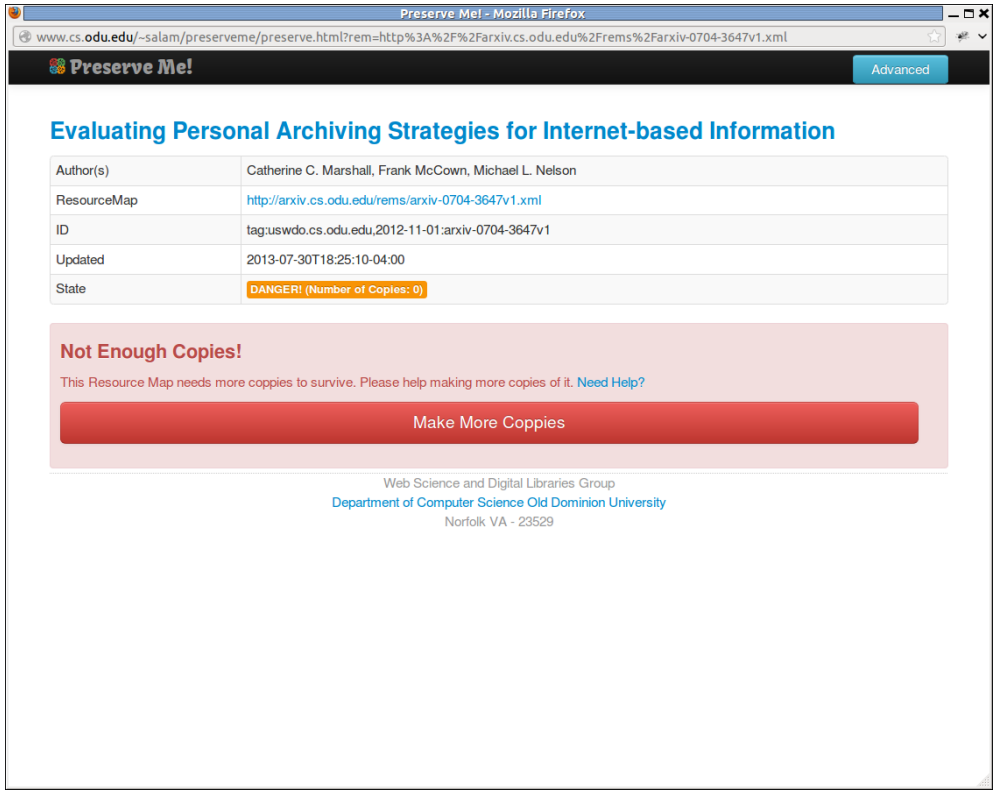
A USW instrumented splash page




```
...
<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/wsdl/uswdo/work/preserveme.js"></script>
...
```

USW algorithm popup

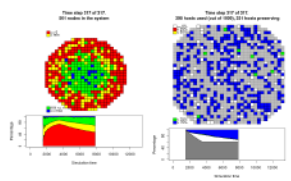
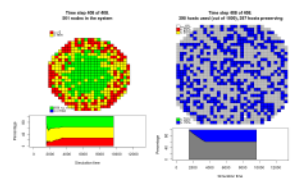
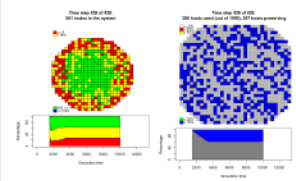
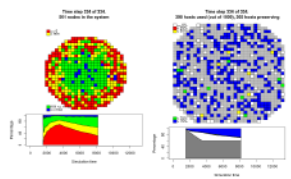
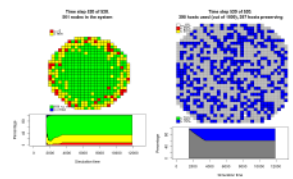
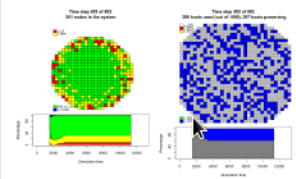
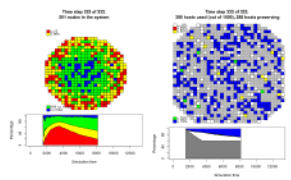
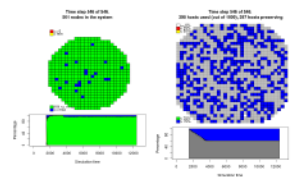
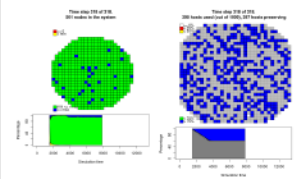
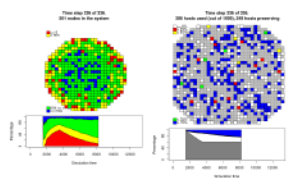
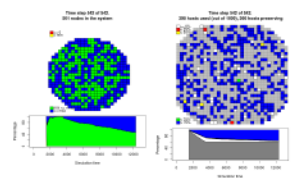
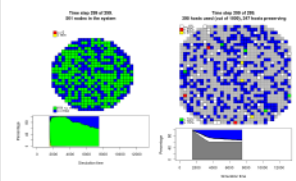
- Written in JavaScript
- Relies on domain services
  - Copy -> creates copy of a WO
  - Edit -> update own REM
- Uses communications mechanism based on Sawood Alam's master's thesis



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

Named condition	Least aggressive	Moderately aggressive	Most aggressive
Famine			
Boundary Low			
Straddle			
Boundary High			
Feast