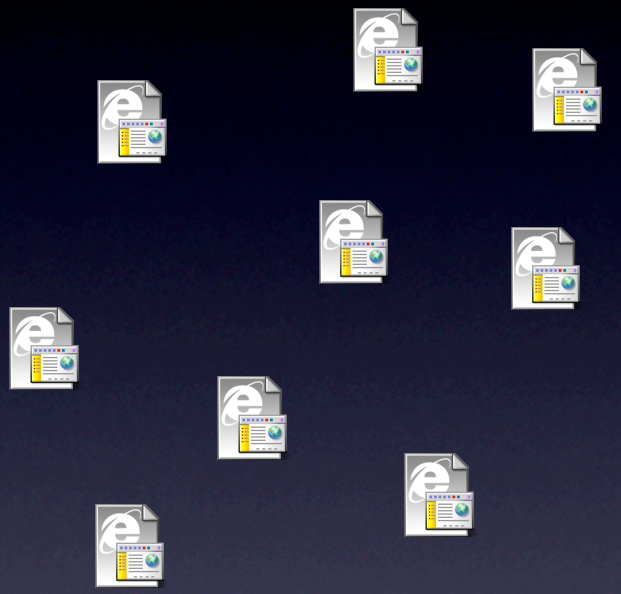
Systems challenges in online social media

Alan Mislove

College of Computer and Information Science
Northeastern University

December 6th, 2011, Top-IX Conference
Turin, Italy

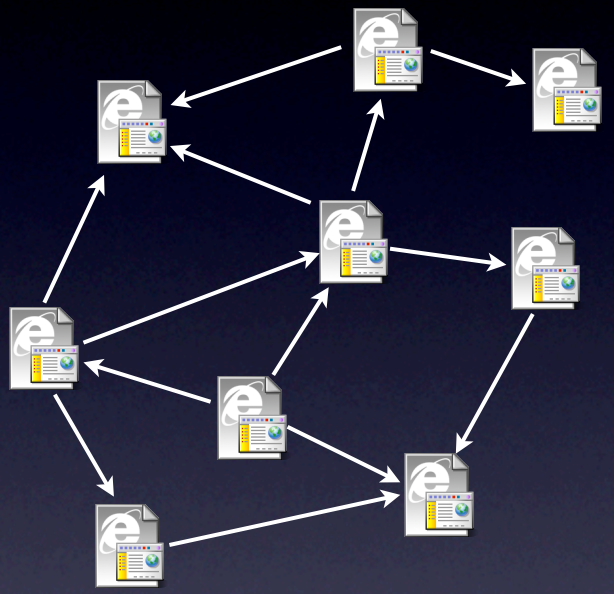
A new way of organizing information



Web

Online social media

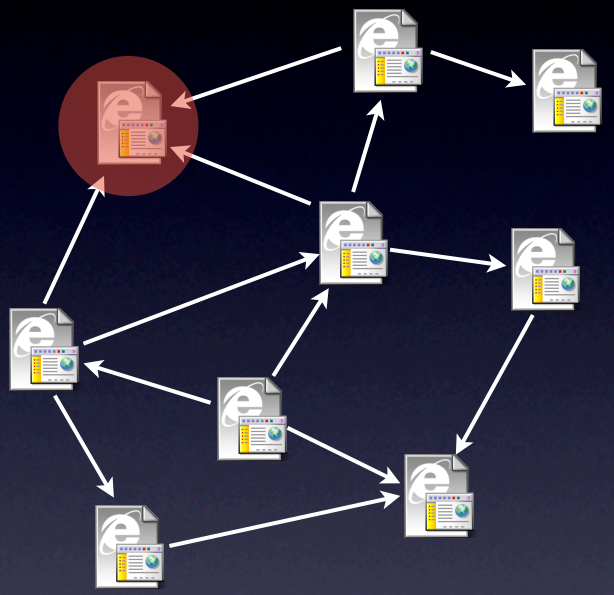
A new way of organizing information



Web

Online social media

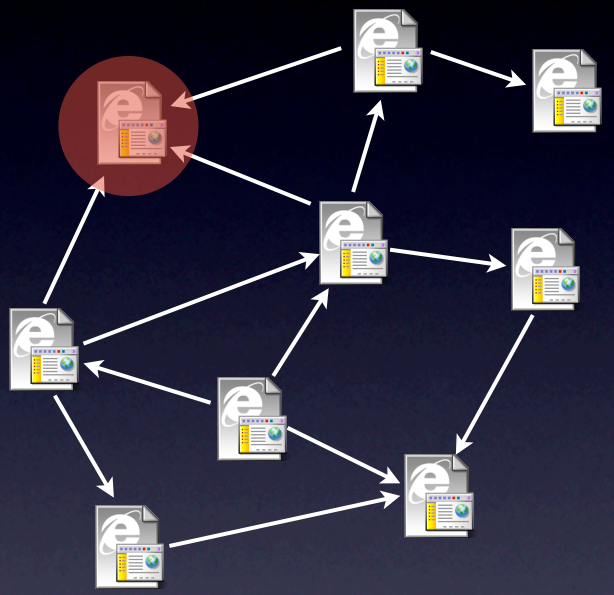
A new way of organizing information



Web

Online social media

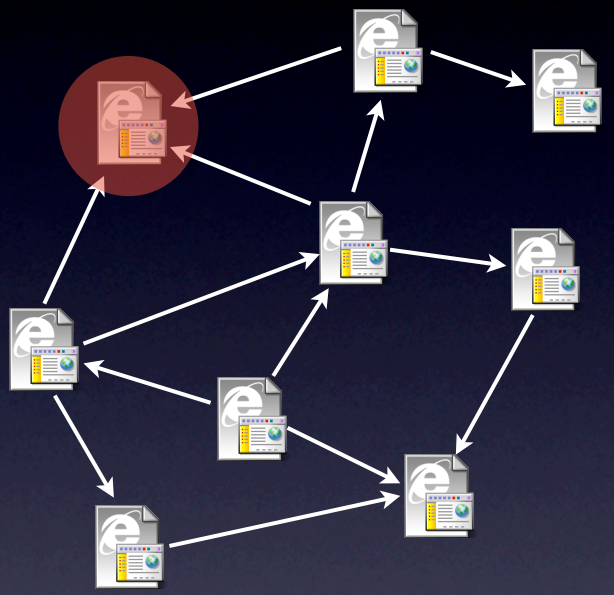
A new way of organizing information



Web

Online social media

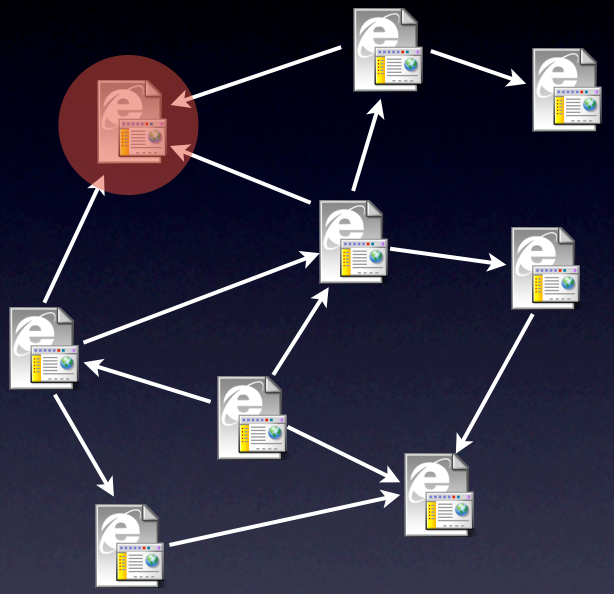
A new way of organizing information



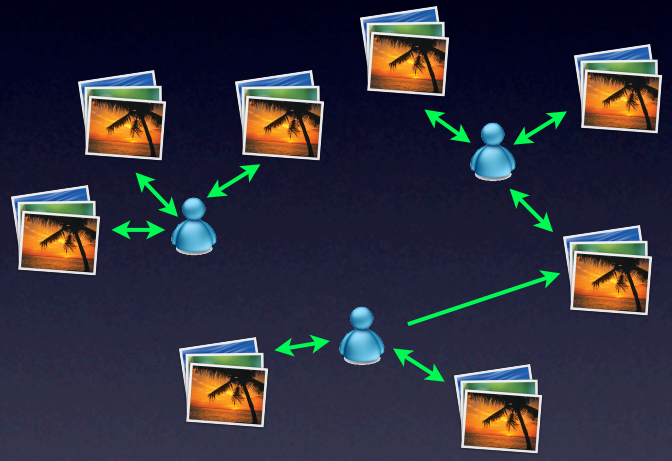
Web

Online social media

A new way of organizing information

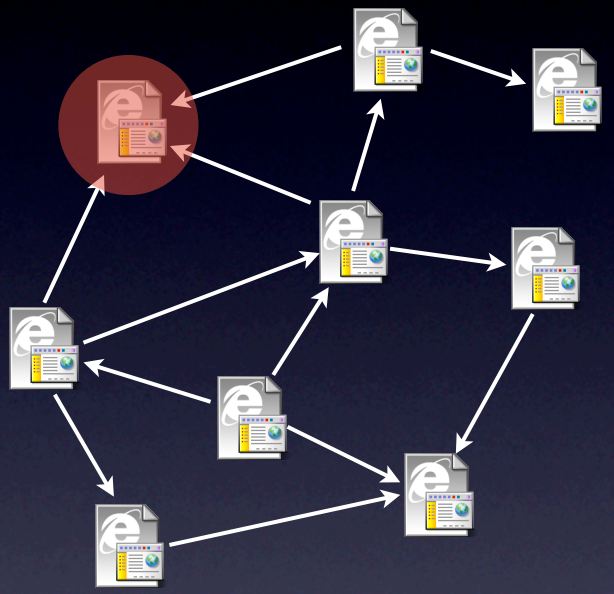


Web

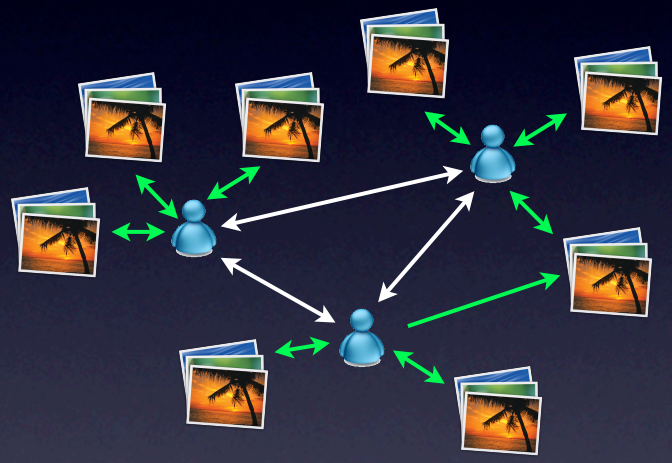


Online social media

A new way of organizing information



Web



Online social media

My group's research

Leveraging social networks results in "better" systems

Due to the increasing integration of systems and social networks

Step 1: Understand social network properties

Step 2: Build systems and algorithms to address challenges

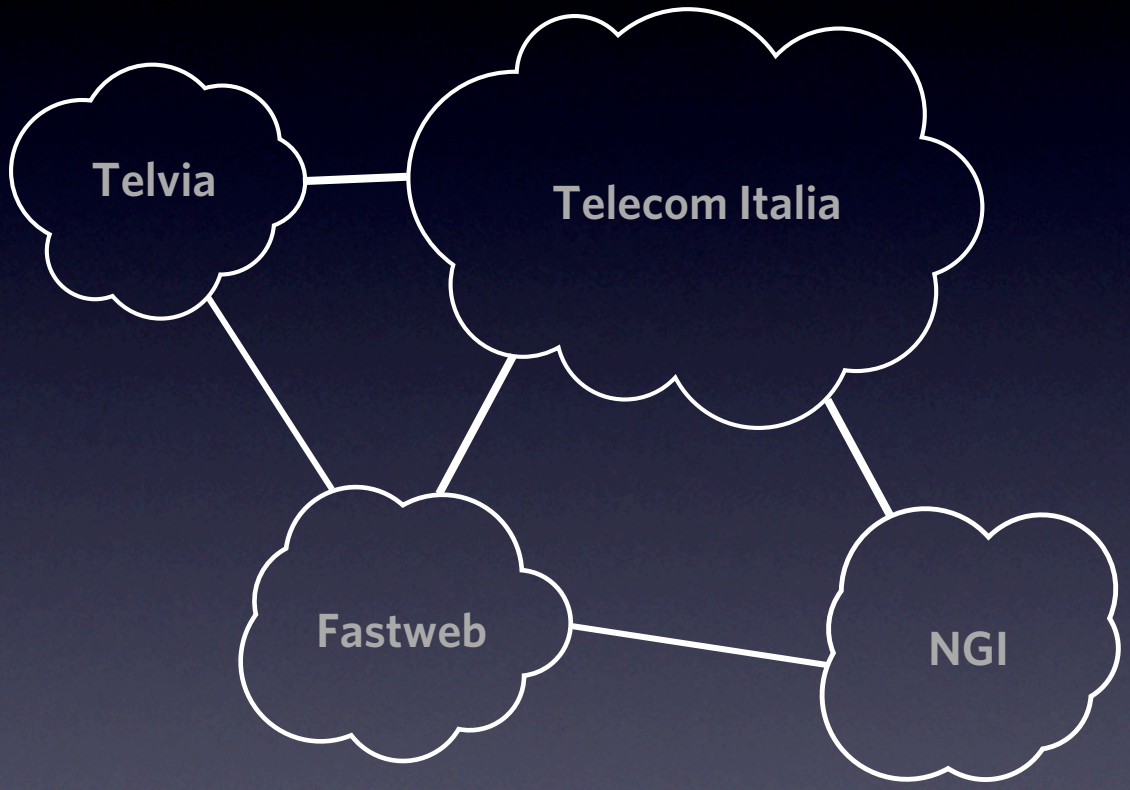
My group's research is motivated by *trends*

Will give two examples today

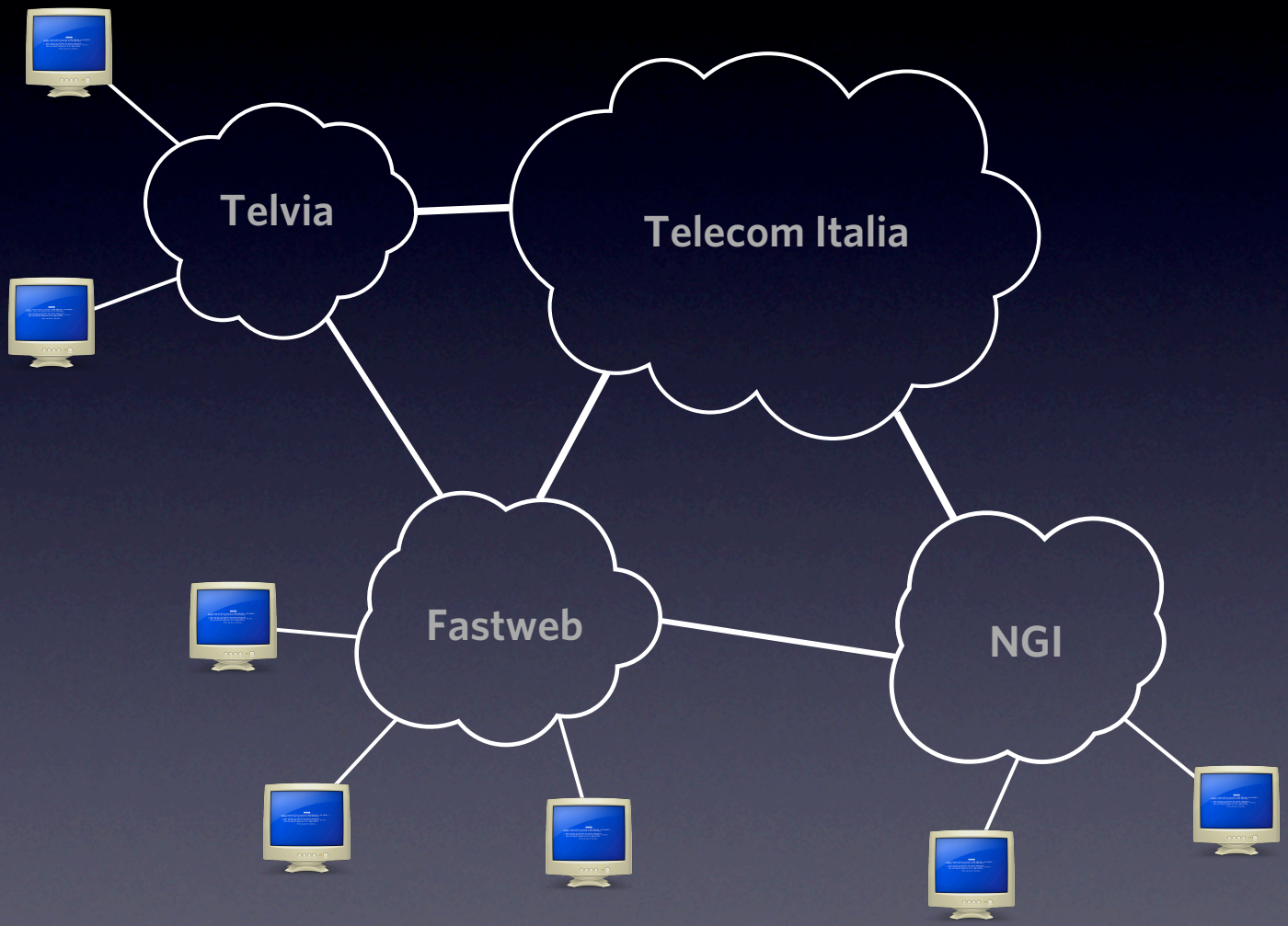
Trend 1:

Changing patterns of content creation + exchange

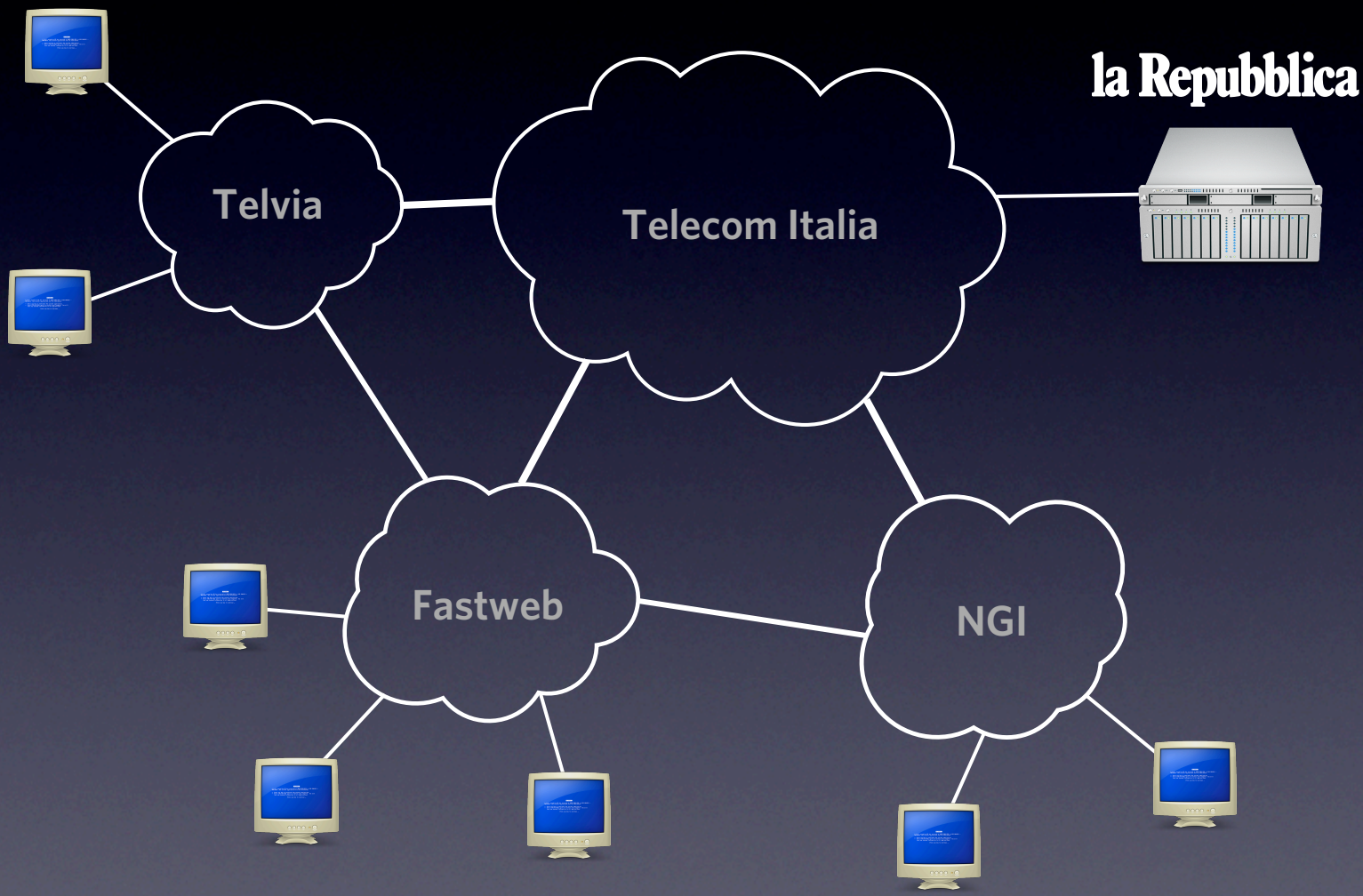
Pre-2005 Web (a.k.a. Web 1.0)



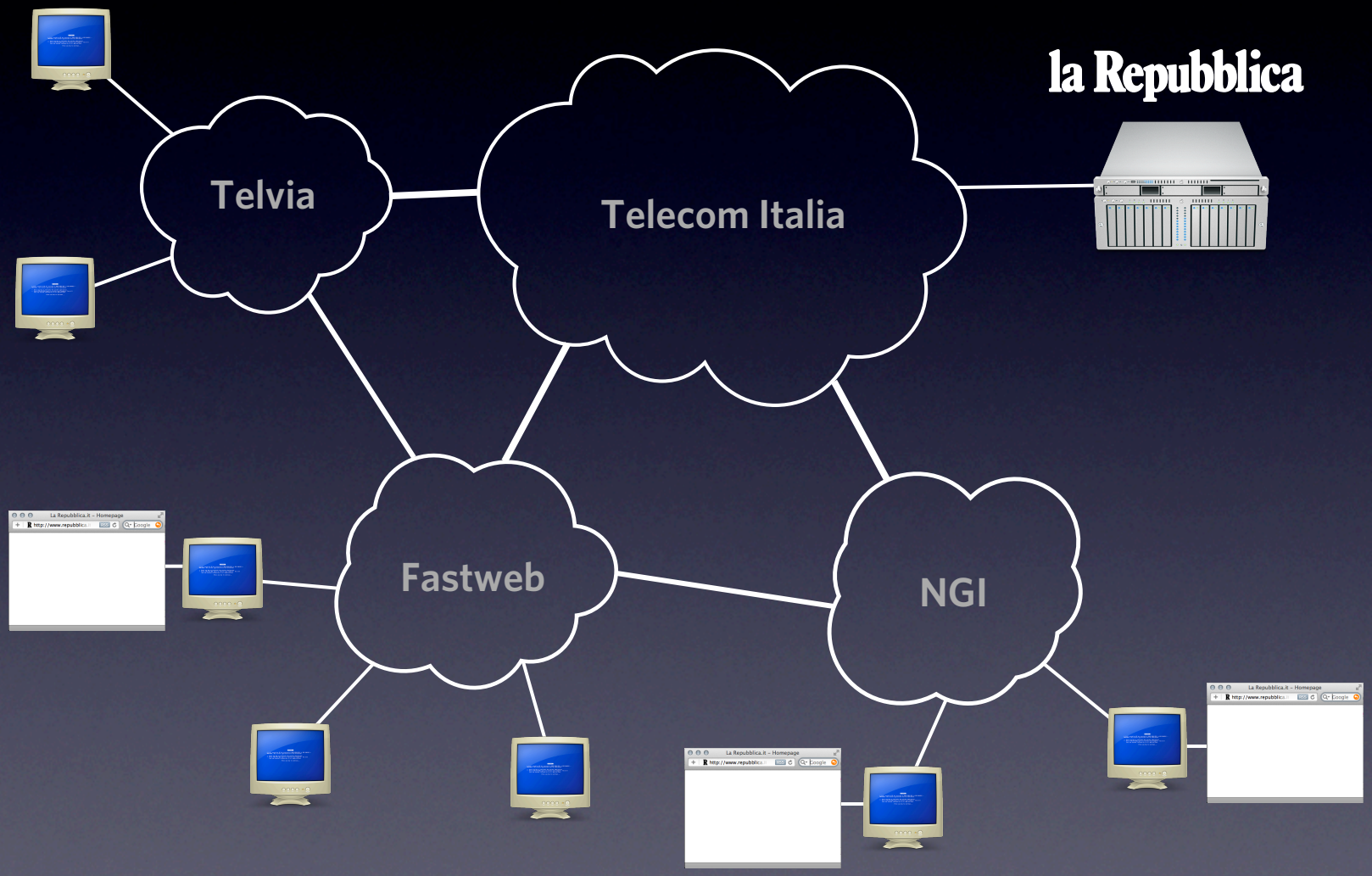
Pre-2005 Web (a.k.a. Web 1.0)



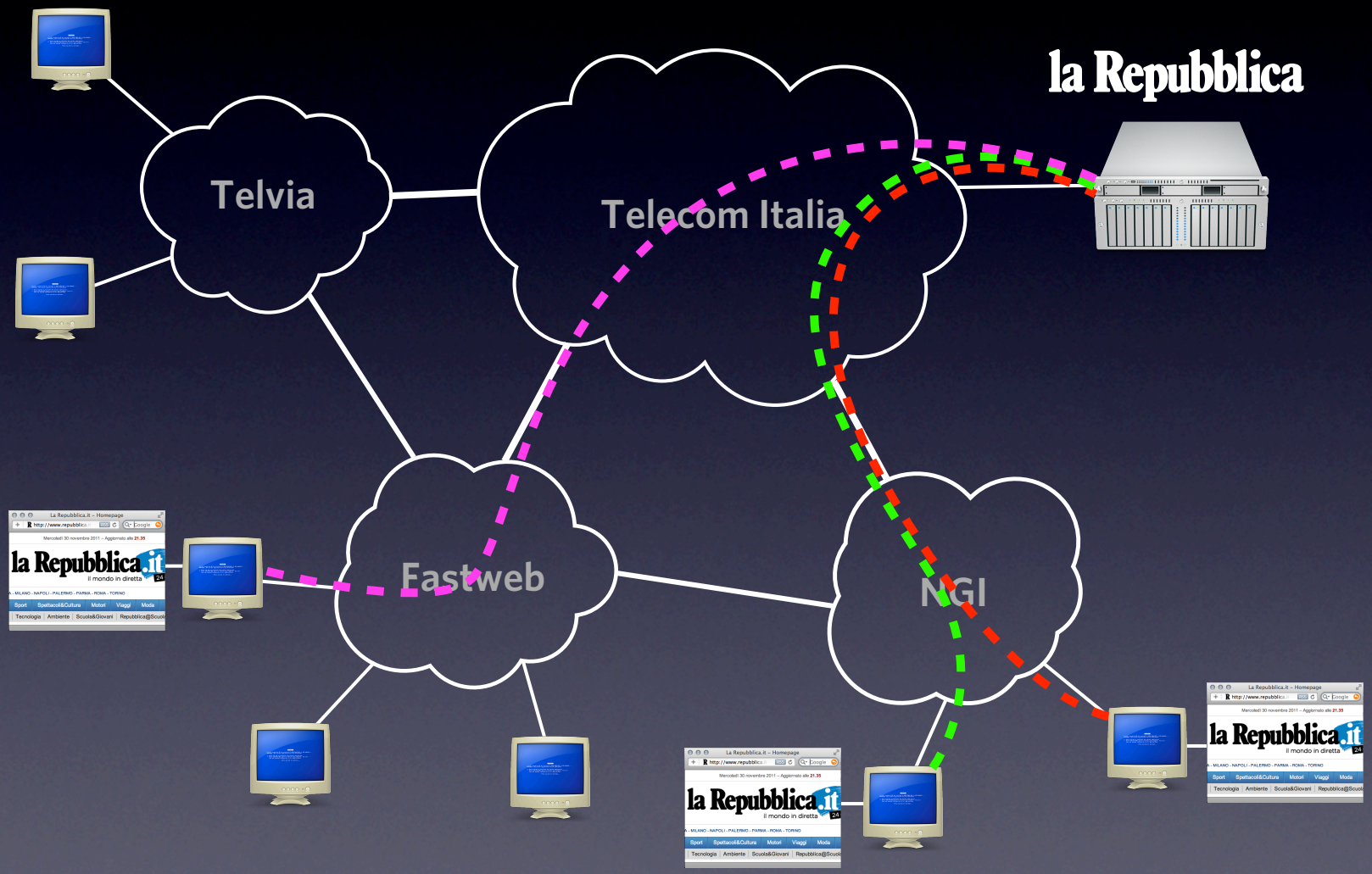
Pre-2005 Web (a.k.a. Web 1.0)



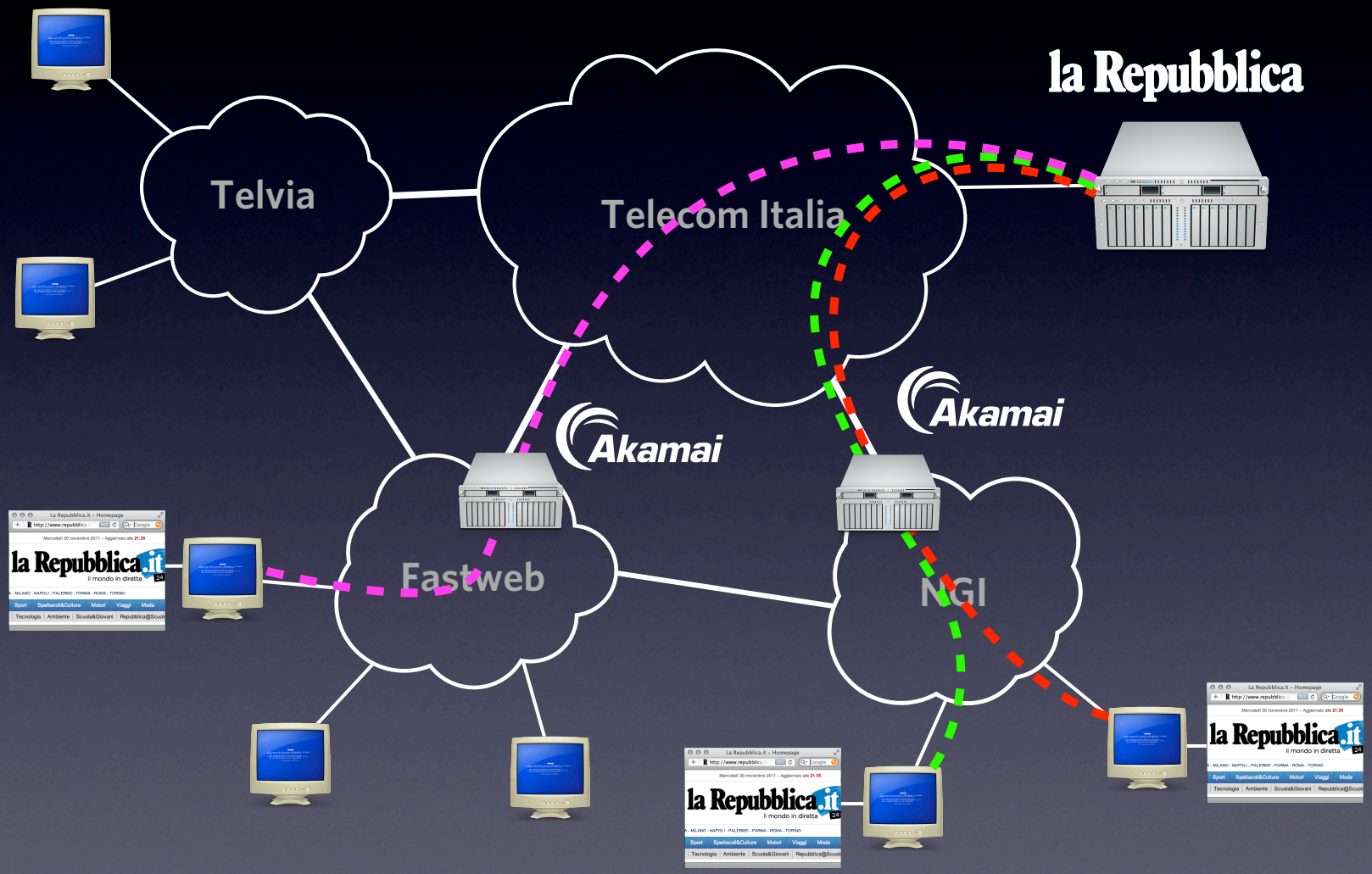
Pre-2005 Web (a.k.a. Web 1.0)



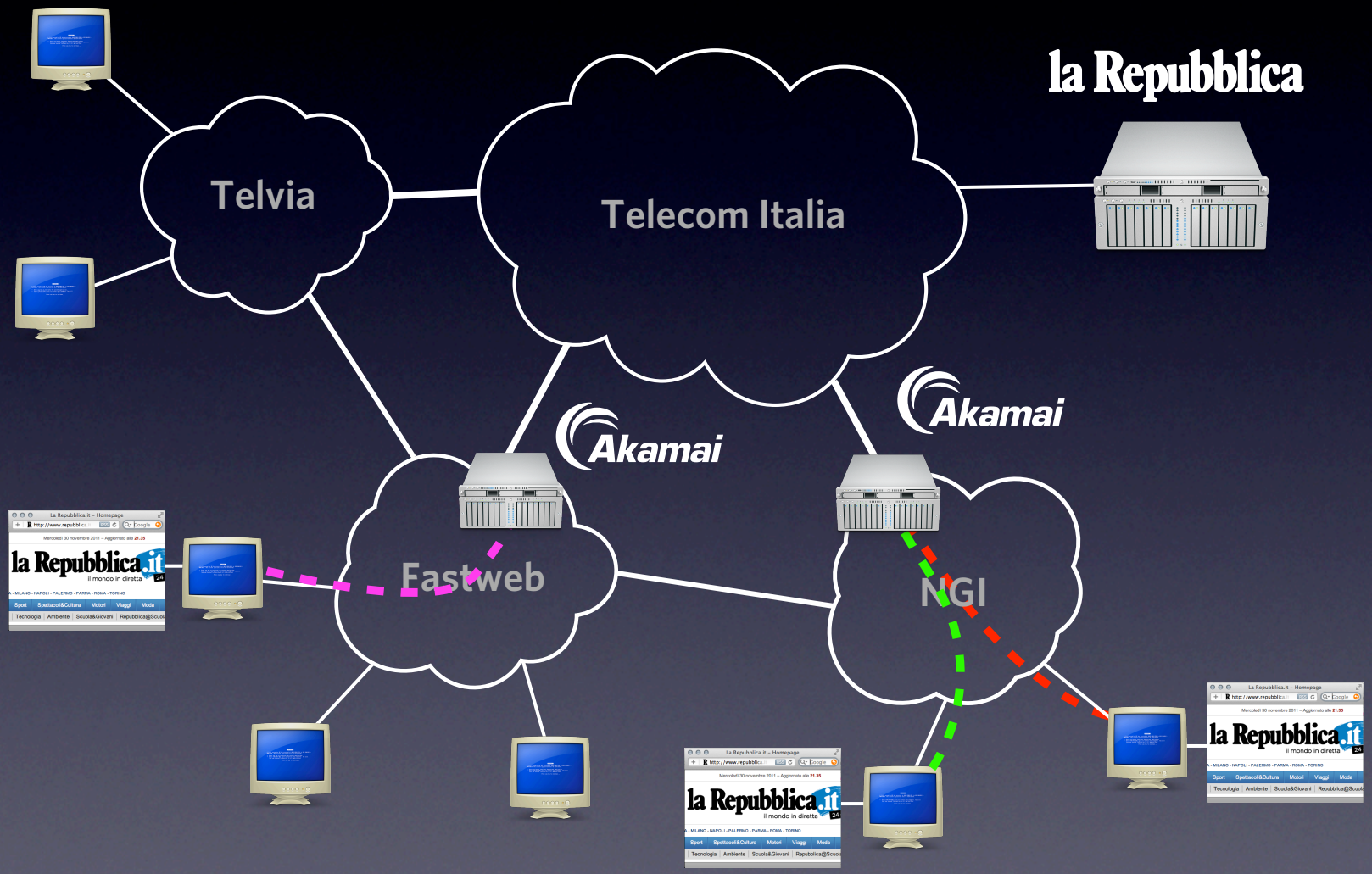
Pre-2005 Web (a.k.a. Web 1.0)



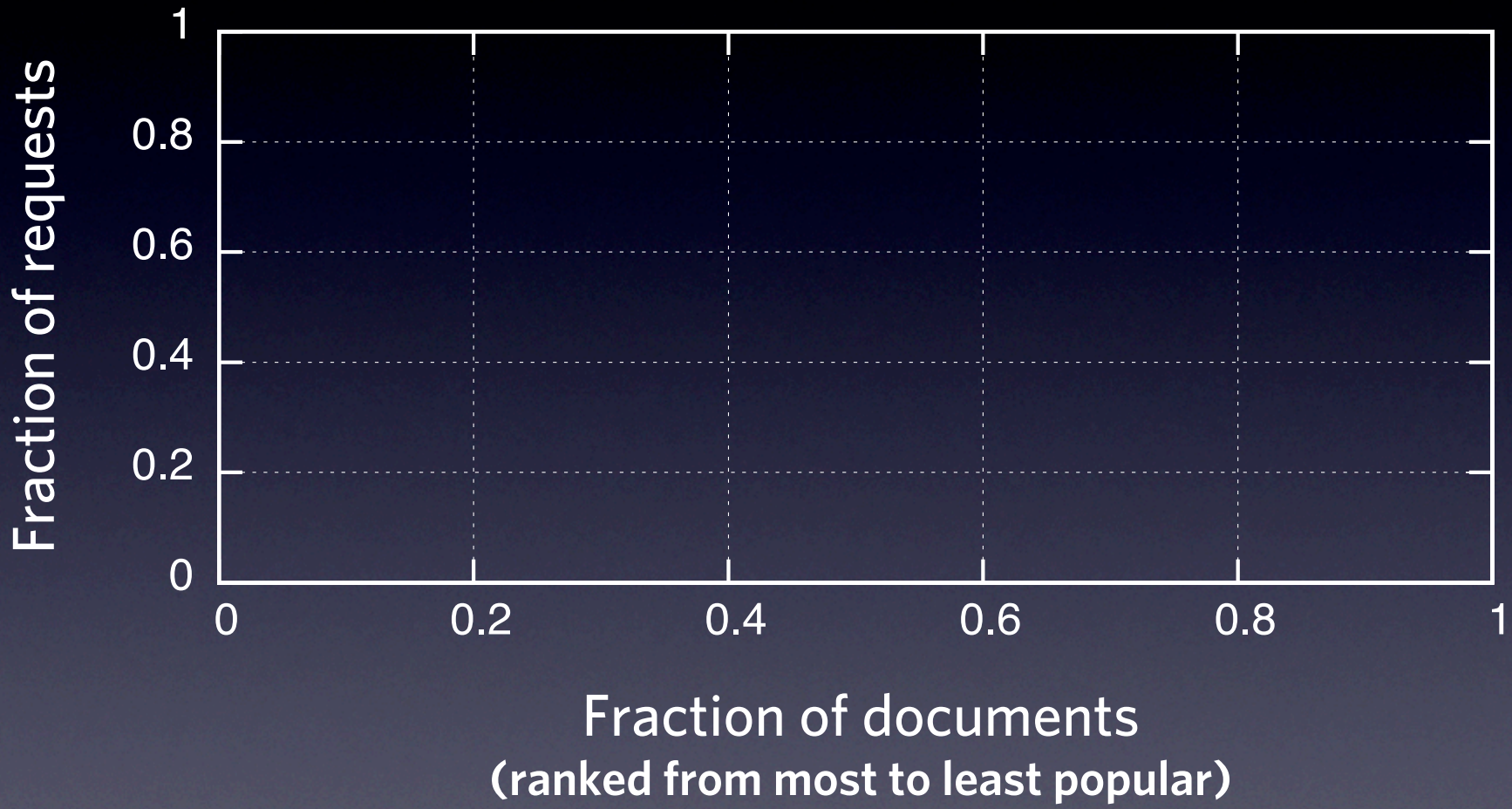
Pre-2005 Web (a.k.a. Web 1.0)



Pre-2005 Web (a.k.a. Web 1.0)

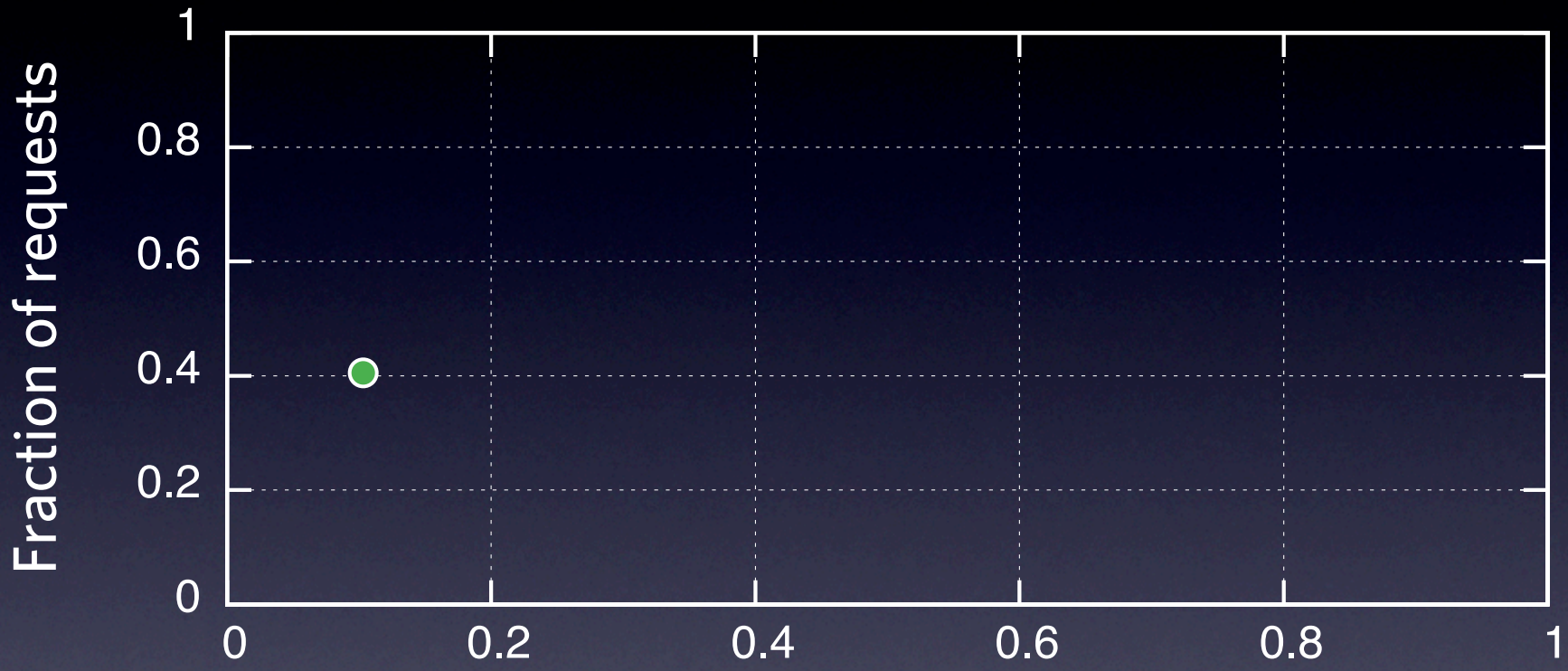


Change 1: Content popularity



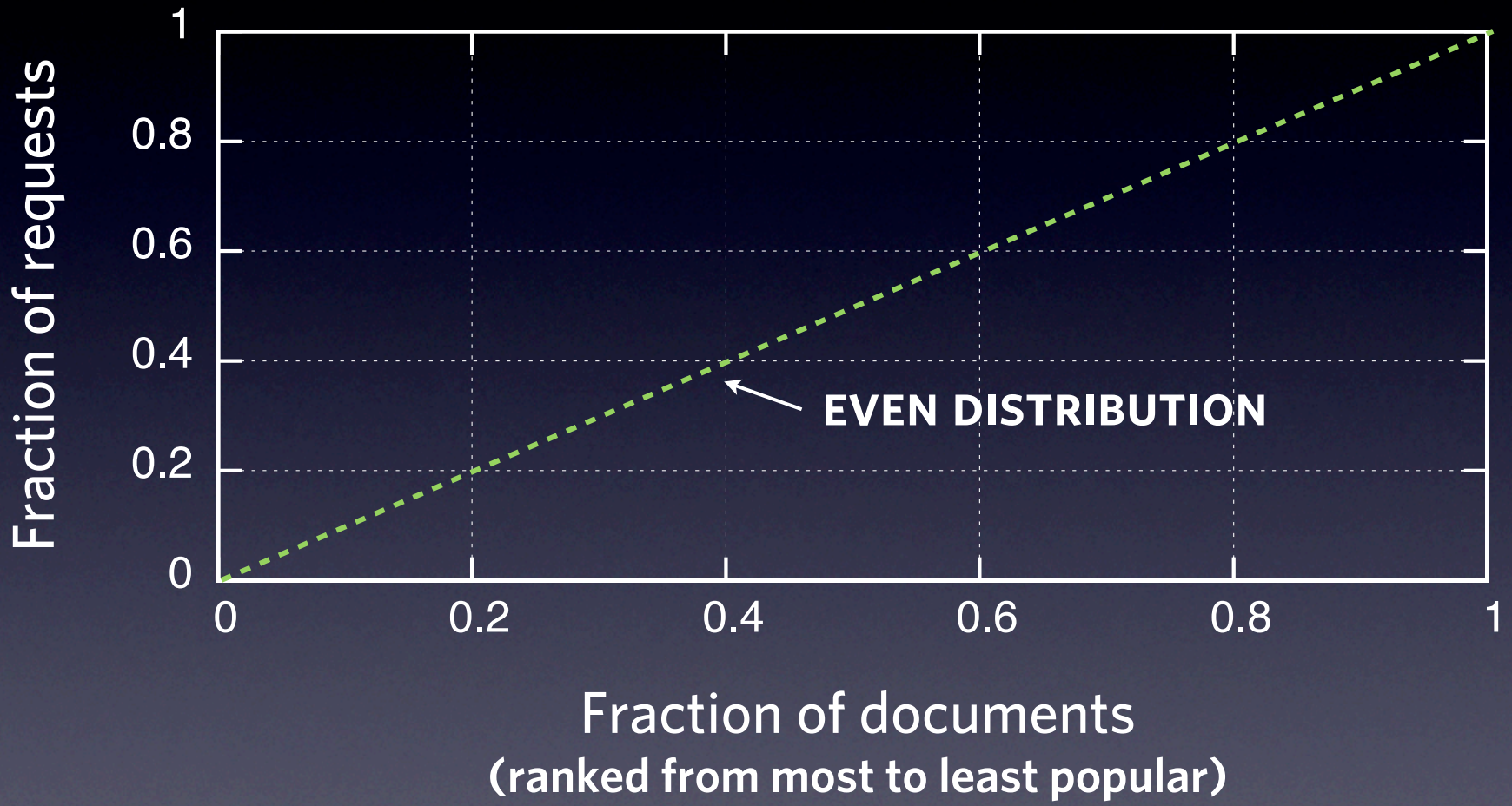
[1] Breslau et al., INFOCOM, 1999, [2] Mislove et al., WSDM, 2010

Change 1: Content popularity



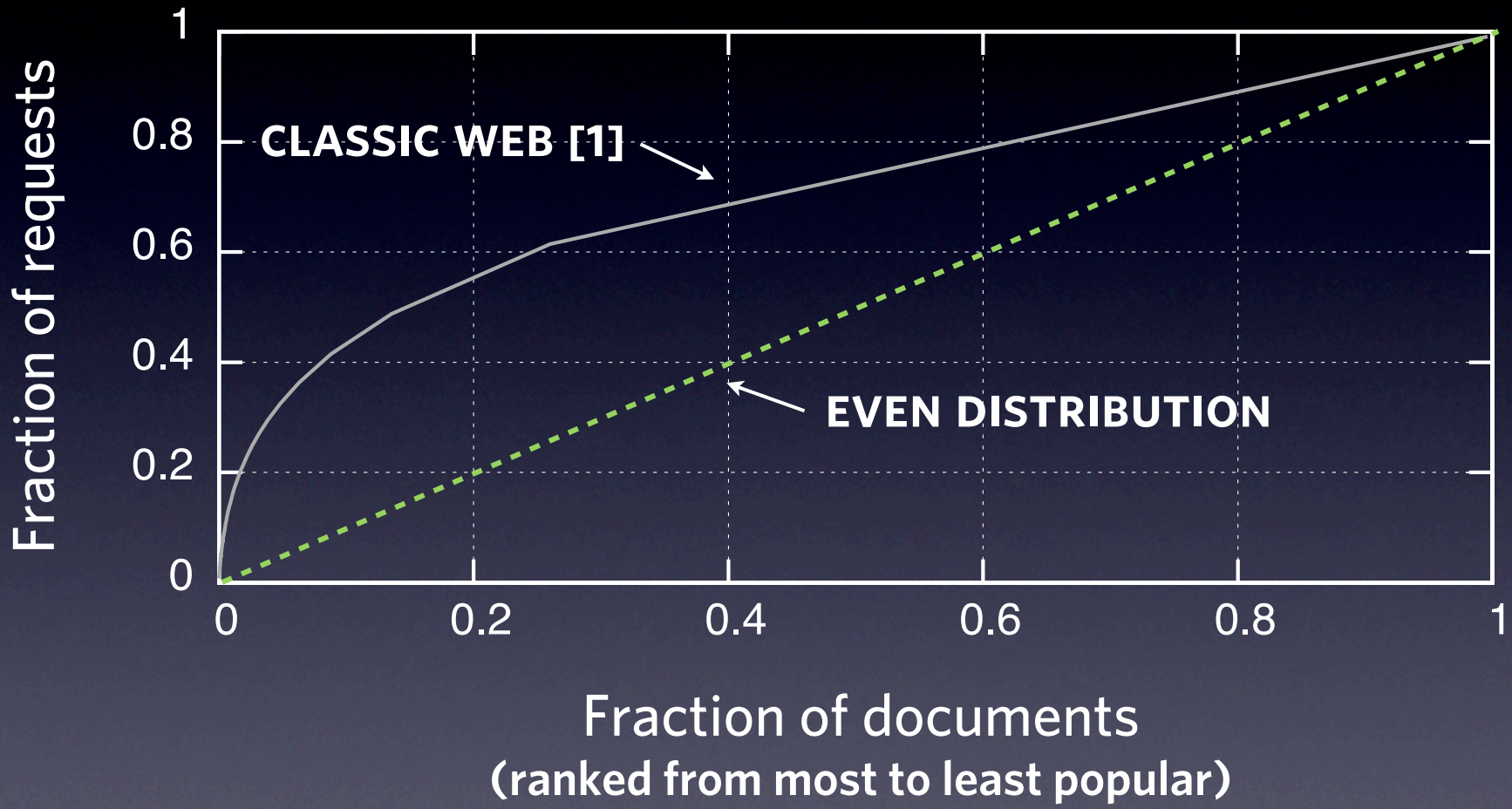
[1] Breslau et al., INFOCOM, 1999, [2] Mislove et al., WSDM, 2010

Change 1: Content popularity



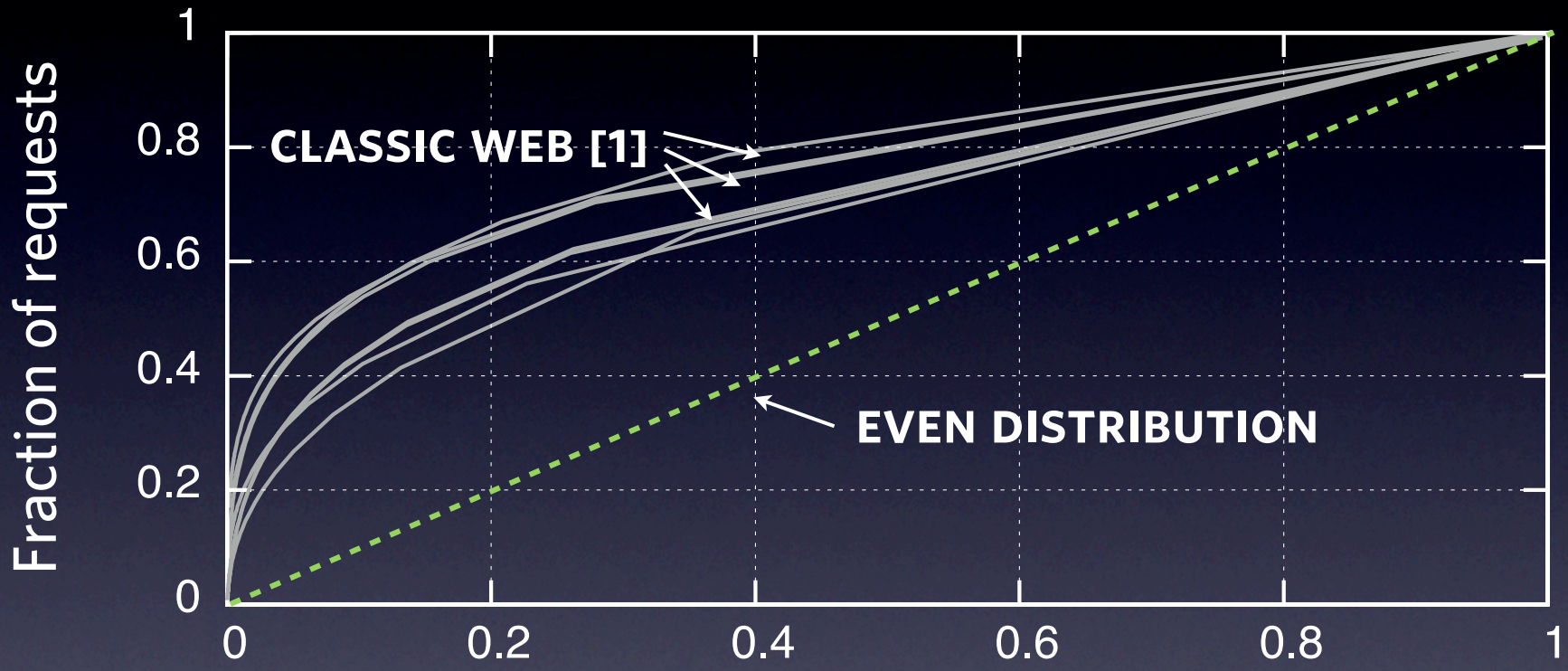
[1] Breslau et al., INFOCOM, 1999, [2] Mislove et al., WSDM, 2010

Change 1: Content popularity



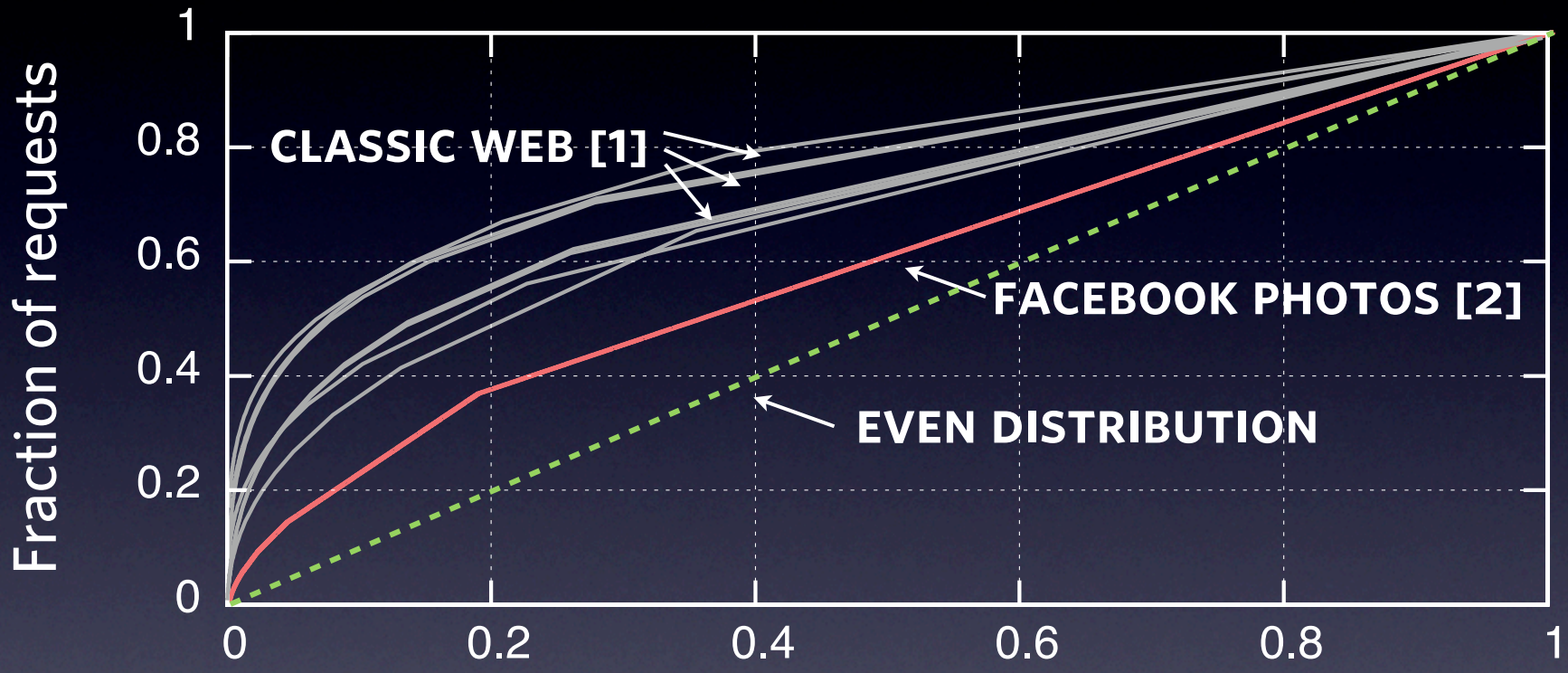
[1] Breslau et al., INFOCOM, 1999, [2] Mislove et al., WSDM, 2010

Change 1: Content popularity



[1] Breslau et al., INFOCOM, 1999, [2] Mislove et al., WSDM, 2010

Change 1: Content popularity



[1] Breslau et al., INFOCOM, 1999, [2] Mislove et al., WSDM, 2010

Implication: Caches less effective

Popularity distribution much more even

- Objects have more narrow scope

In classic Web:

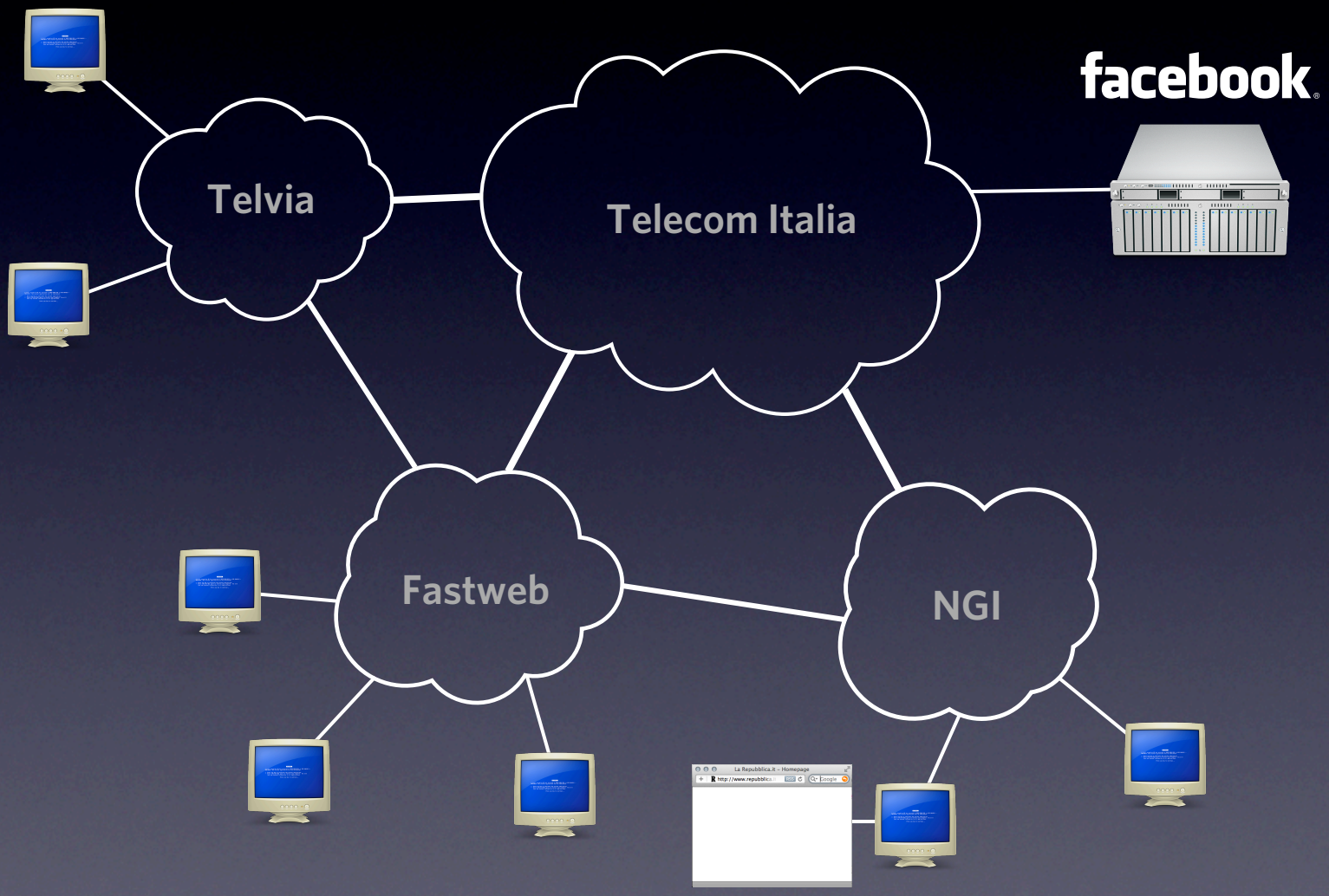
- Caching top 10% serves between 55% [1] and 95% [2] of requests
- Success of CDNs, web caches, …

In online social media:

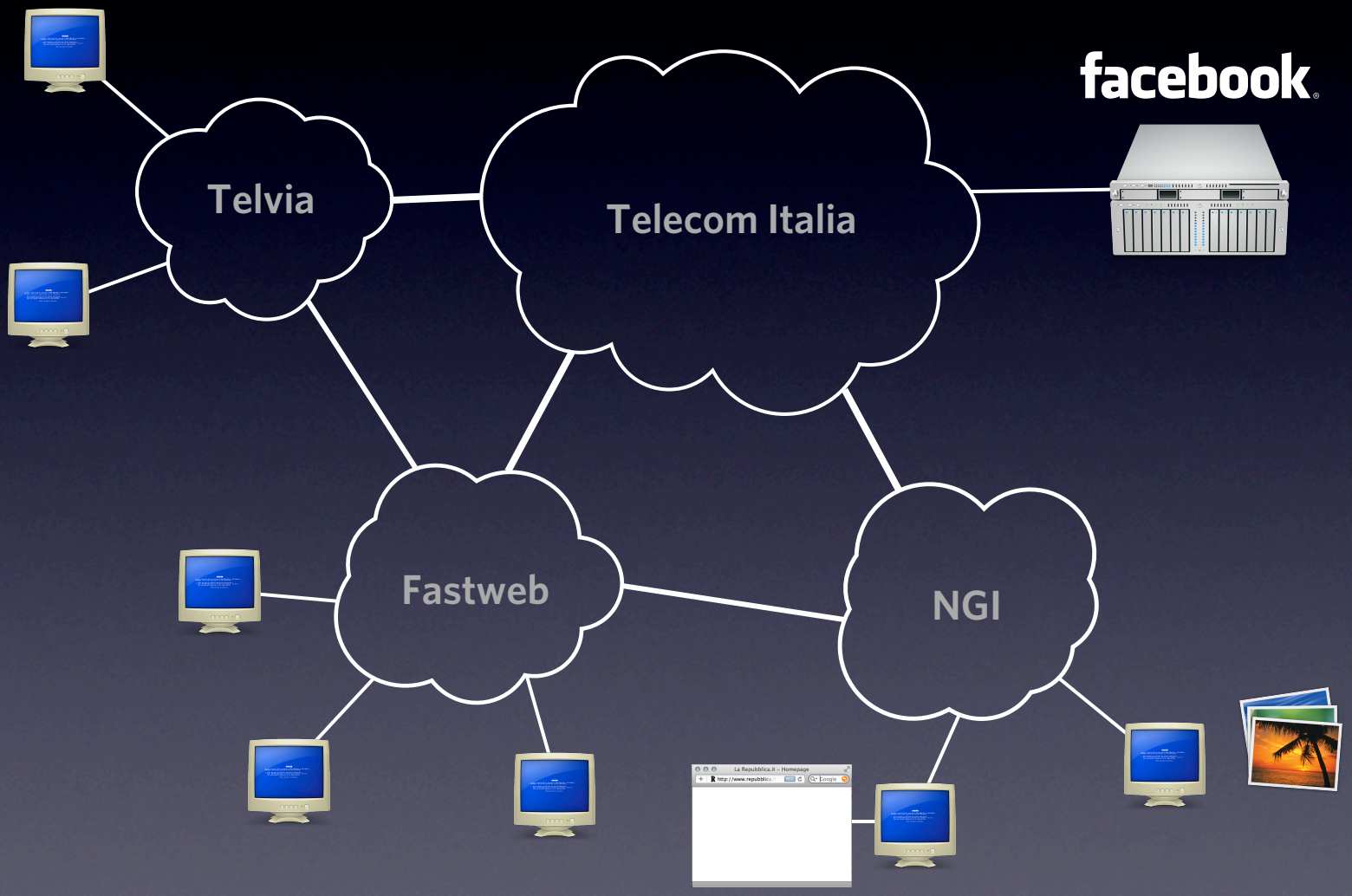
- Caching top 10% would only serve 27% [3] of requests

[1] Breslau et al., INFOCOM, 1999, [2] Arlitt et al. IEEE Network, 2000, [3] Mislove et al., WSDM, 2010

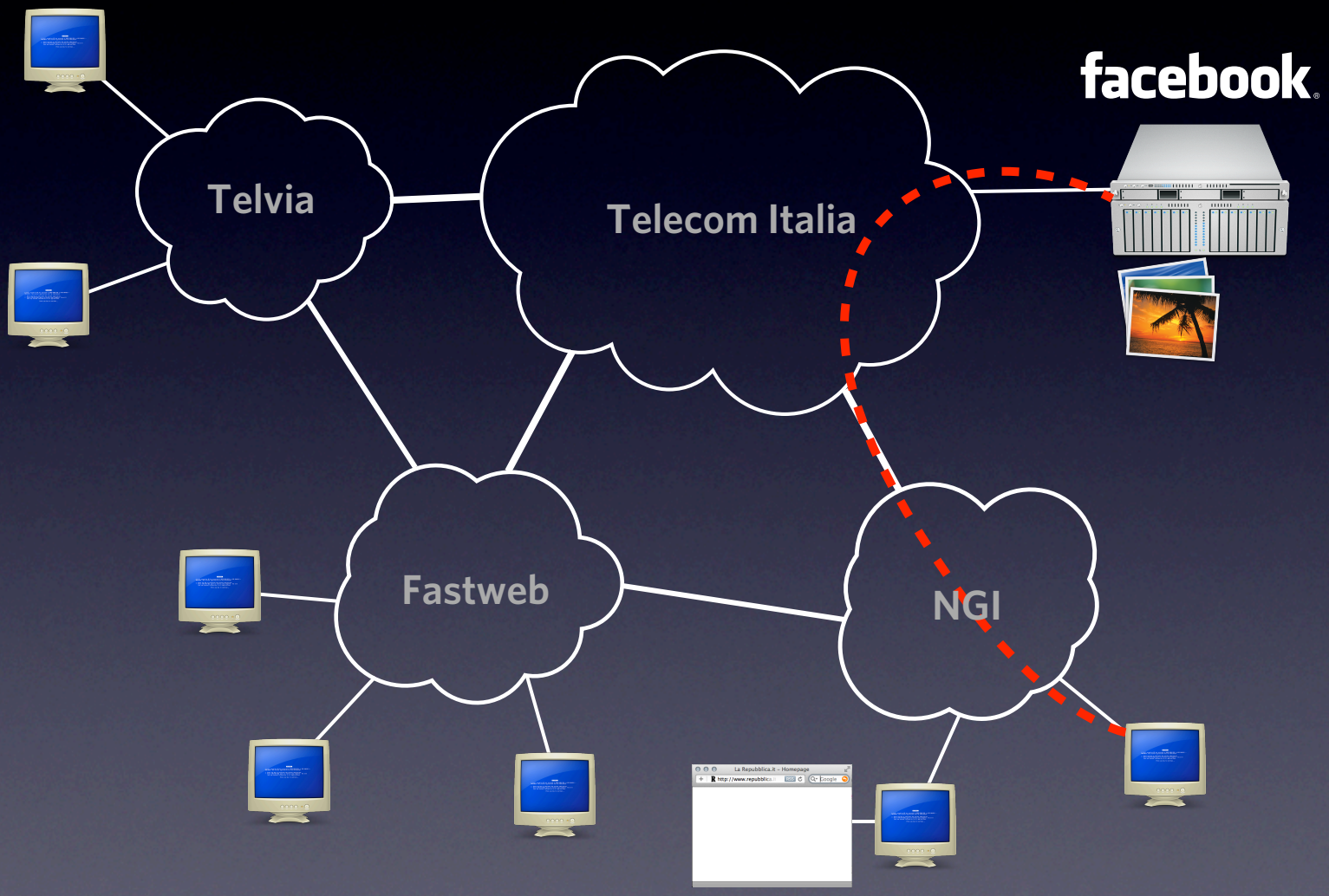
Change 2: Content generation



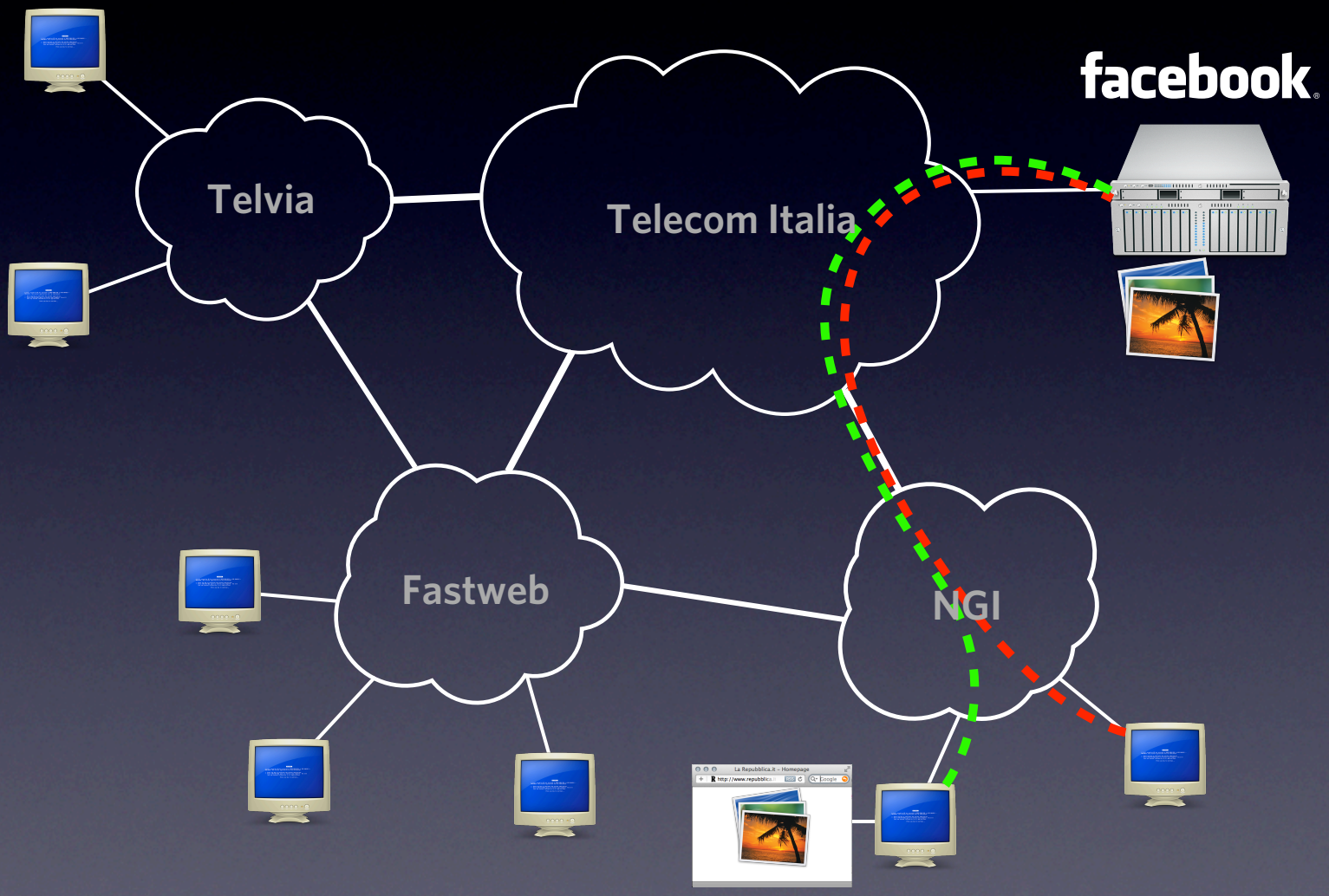
Change 2: Content generation



Change 2: Content generation



Change 2: Content generation



Implication: Workload change

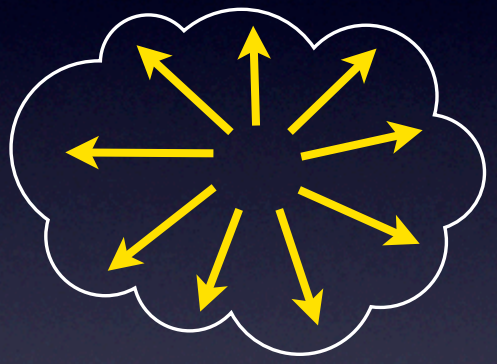
Significant content creation at network's edge
- Ease of digital content creation (photos, video)
- Ubiquity of Internet access (cell phone, iPad)

In classic Web:
- Workload was "center-to-edge"
- Caching, CDNs take load off origin server

In online social media:
- Workload is "edge-to-edge"
- Significant geographic locality

Implication: Workload change

Significant content creation at network's edge
- Ease of digital content creation (photos, video)
- Ubiquity of Internet access (cell phone, iPad)



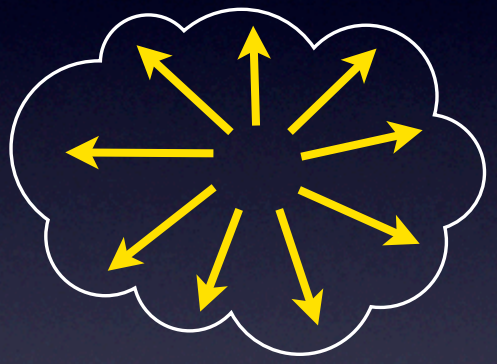
In classic Web:
- Workload was "center-to-edge"
- Caching, CDNs take load off origin server

In online social media:
- Workload is "edge-to-edge"
- Significant geographic locality

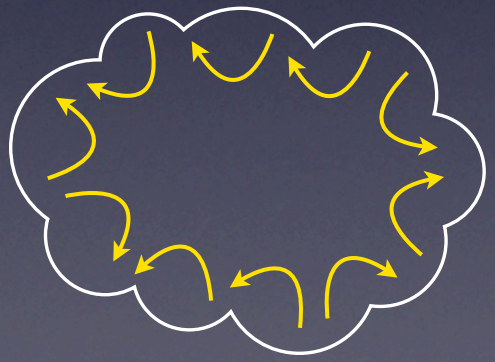
Implication: Workload change

Significant content creation at network's edge
- Ease of digital content creation (photos, video)
- Ubiquity of Internet access (cell phone, iPad)

In classic Web:
- Workload was "center-to-edge"
- Caching, CDNs take load off origin server



In online social media:
- Workload is "edge-to-edge"
- Significant geographic locality



How is OSN content being delivered?

Web 1.0 "centralized" architectures dominate
- Akamai, Limelight, Clearway, …
- Facebook serves much of its own content

Mismatch between infrastructure, workload

Workload is naturally decentralized
- Every Facebook upload goes via CA

Can we build a workload-matching distribution system?
- Avoid unnecessary, expensive transfers

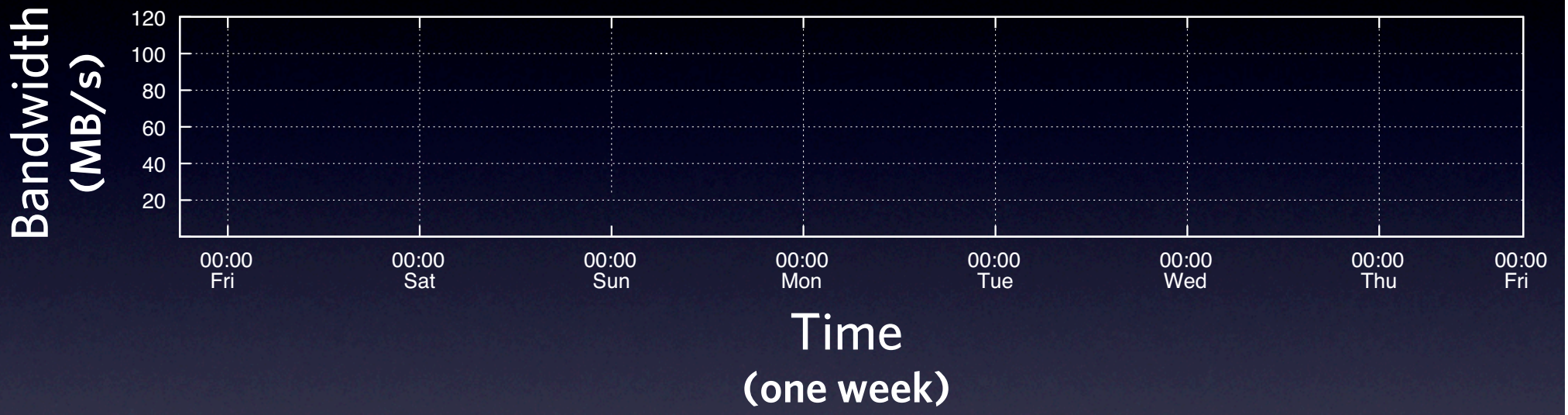
WebCloud: Decentralized delivery

First step towards decentralized Web content delivery
- Challenge: Web doesn't support decentralization
- Browsers distinct from Web servers

Use novel techniques to allow browser to serve content
- No client-side changes
- Users help serve content they upload
- Result: Scalable, workload-matching architecture

Don't have time for technical discussion
- Built, deployed prototype

WebCloud applied to real-world site



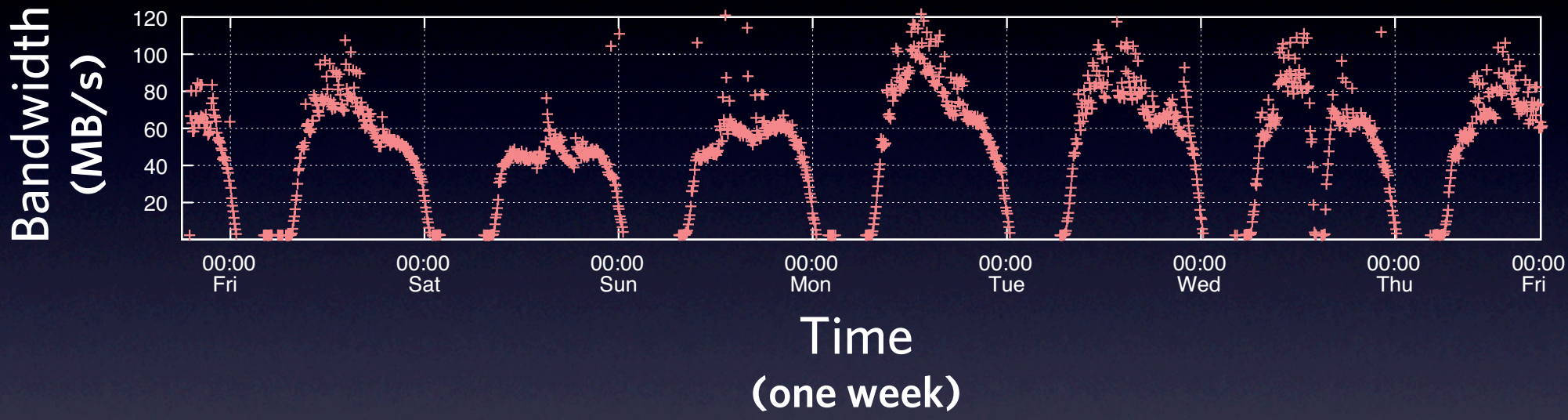
Top-50 U.S. web site

- Simulation based on Akamai logs

Would dramatically reduce bandwidth required

- Savings for both site and ISP

WebCloud applied to real-world site



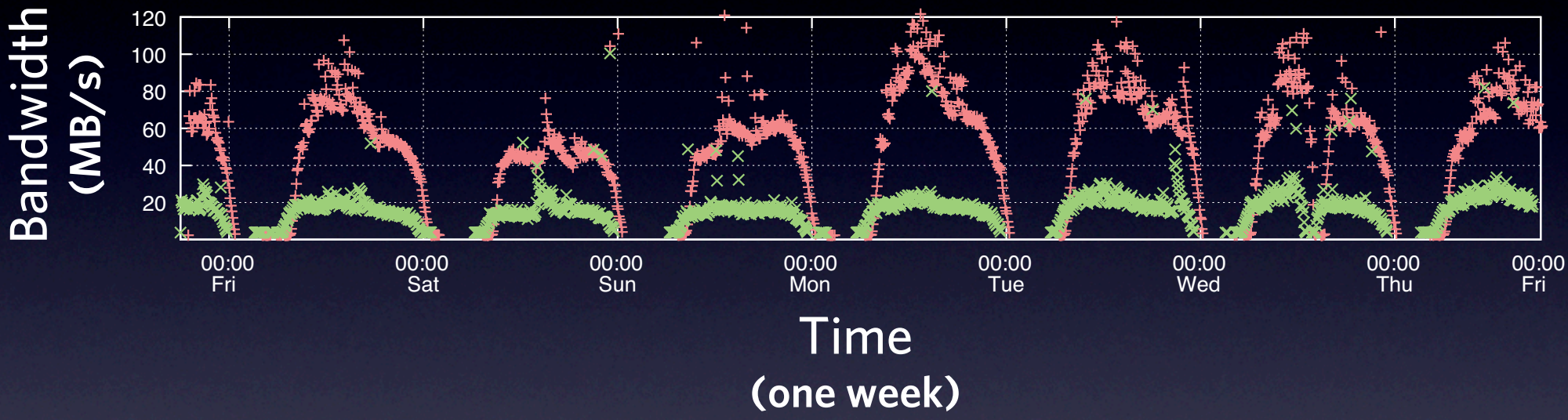
Top-50 U.S. web site

- Simulation based on Akamai logs

Would dramatically reduce bandwidth required

- Savings for both site and ISP

WebCloud applied to real-world site



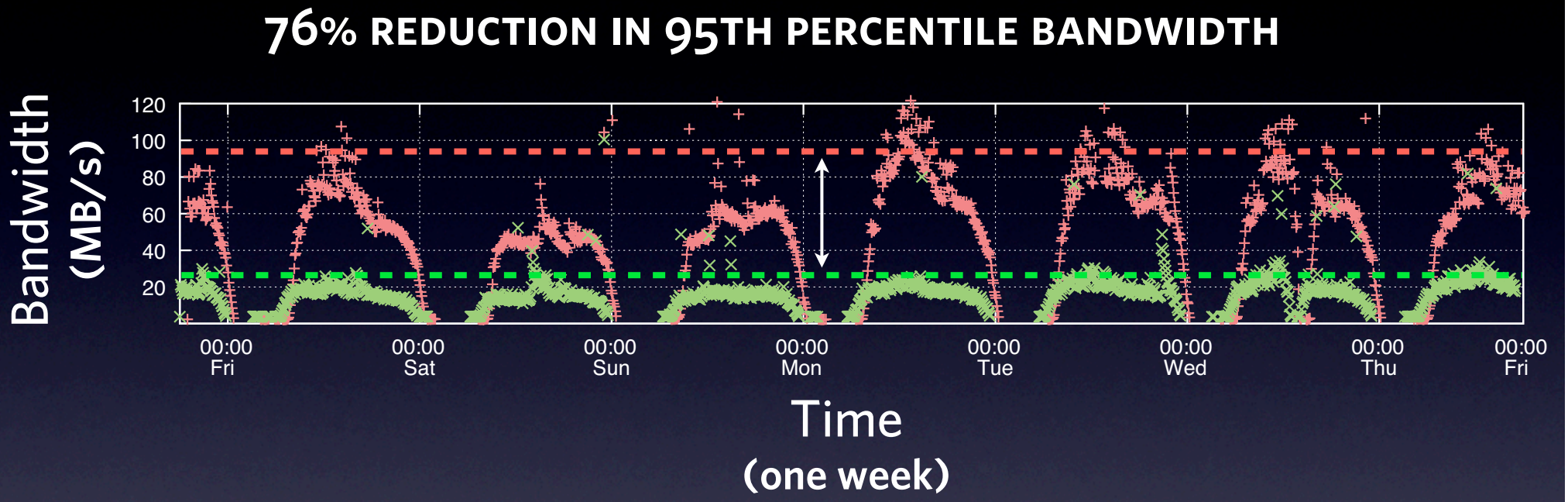
Top-50 U.S. web site

- Simulation based on Akamai logs

Would dramatically reduce bandwidth required

- Savings for both site and ISP

WebCloud applied to real-world site



Top-50 U.S. web site

- Simulation based on Akamai logs

Would dramatically reduce bandwidth required

- Savings for both site and ISP

Summary

Beginnings of shift in patterns of content creation + exchange
- Patterns changing from "center to edge" to "edge to edge"
- Less biased popularity distribution

But, still using centralized delivery architectures

WebCloud: Step towards decentralized Web content delivery
- Users help serve content they create
- Implemented using existing browser features; no client changes

Evaluation demonstrated practicality, efficacy

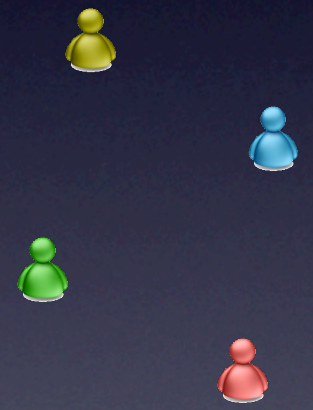
Trend 2:

Changing notion of accounts/identity

User accounts

Account abstraction now ubiquitous
- Represents one or more people in a computer system
- Encapsulates privileges

Traditionally verified by service operators



Trend: Online services with free accounts
- Not verified by operators

Accounts come with privileges
- Send messages (Gmail)
- Upload content (Facebook)
- Vote (Digg)

Sybils

Free accounts with privileges leading to Sybil attacks [IPTPS 2002]

- Single person creates many accounts

Why?

- Natural: Gain extra privileges
- Incentives set up to encourage this

Examples in the wild

- Maze [ICDCS 2007]
- Digg [NSDI 2009]
- TripAdvisor [NYT, 10/2011]
- Facebook, Gmail [me, others]

Sybils

Free accounts with privileges leading to Sybil attacks [IPTPS 2002]

- Single person creates many accounts

Why?

- Natural: Gain extra privileges
- Incentives set up to encourage this

Examples in the wild

- Maze [ICDCS 2007]
- Digg [NSDI 2009]
- TripAdvisor [NYT, 10/2011]
- Facebook, Gmail [me, others]

Example: Online marketplaces

Among most successful Web sites
- eBay alone: $62 B in 2010

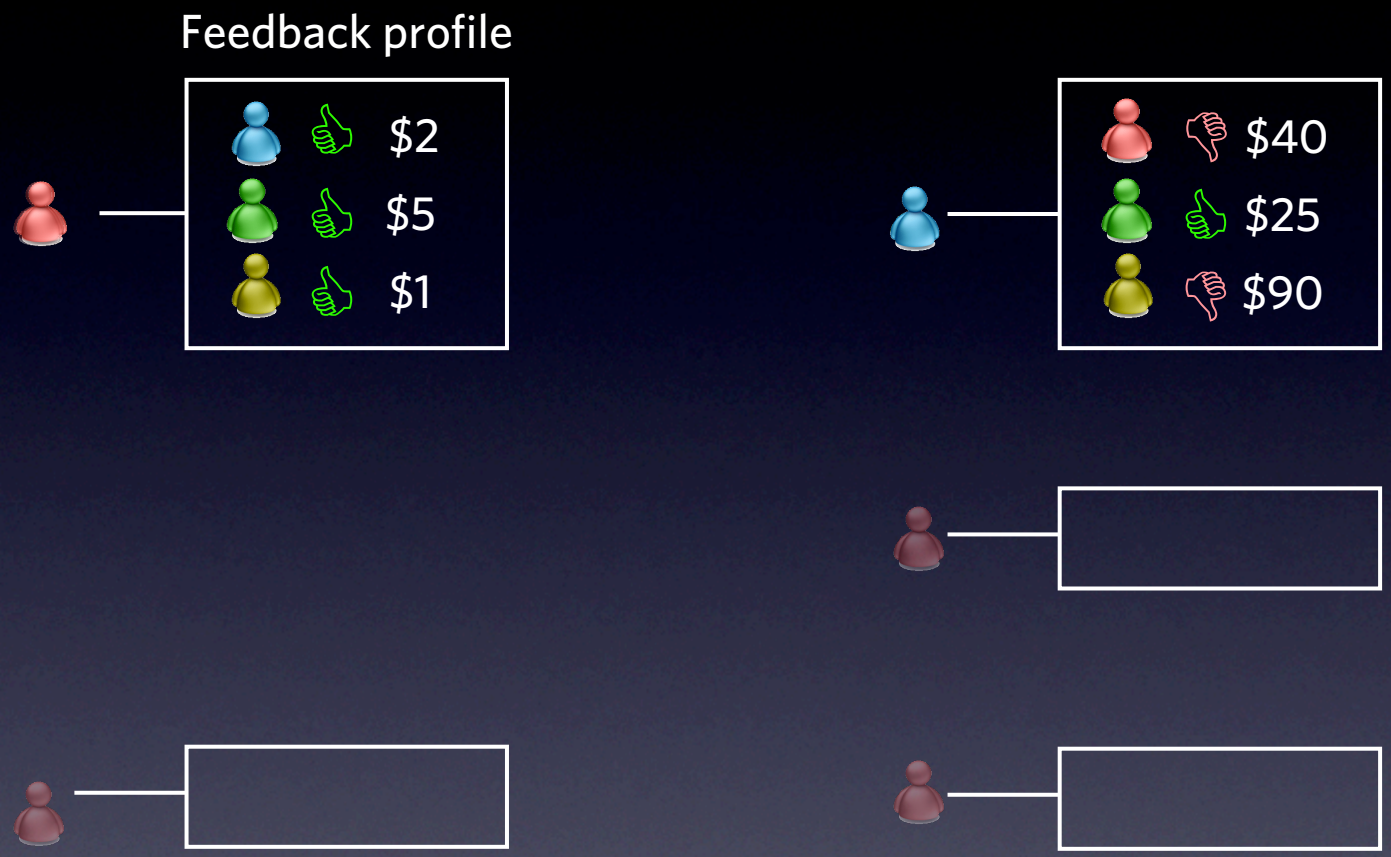
But, known to suffer from fraud

Identities and reputations

Significant monetary losses

- Recent arrest of user who stole $717 κ from 5,000 users
- Used >250 accounts

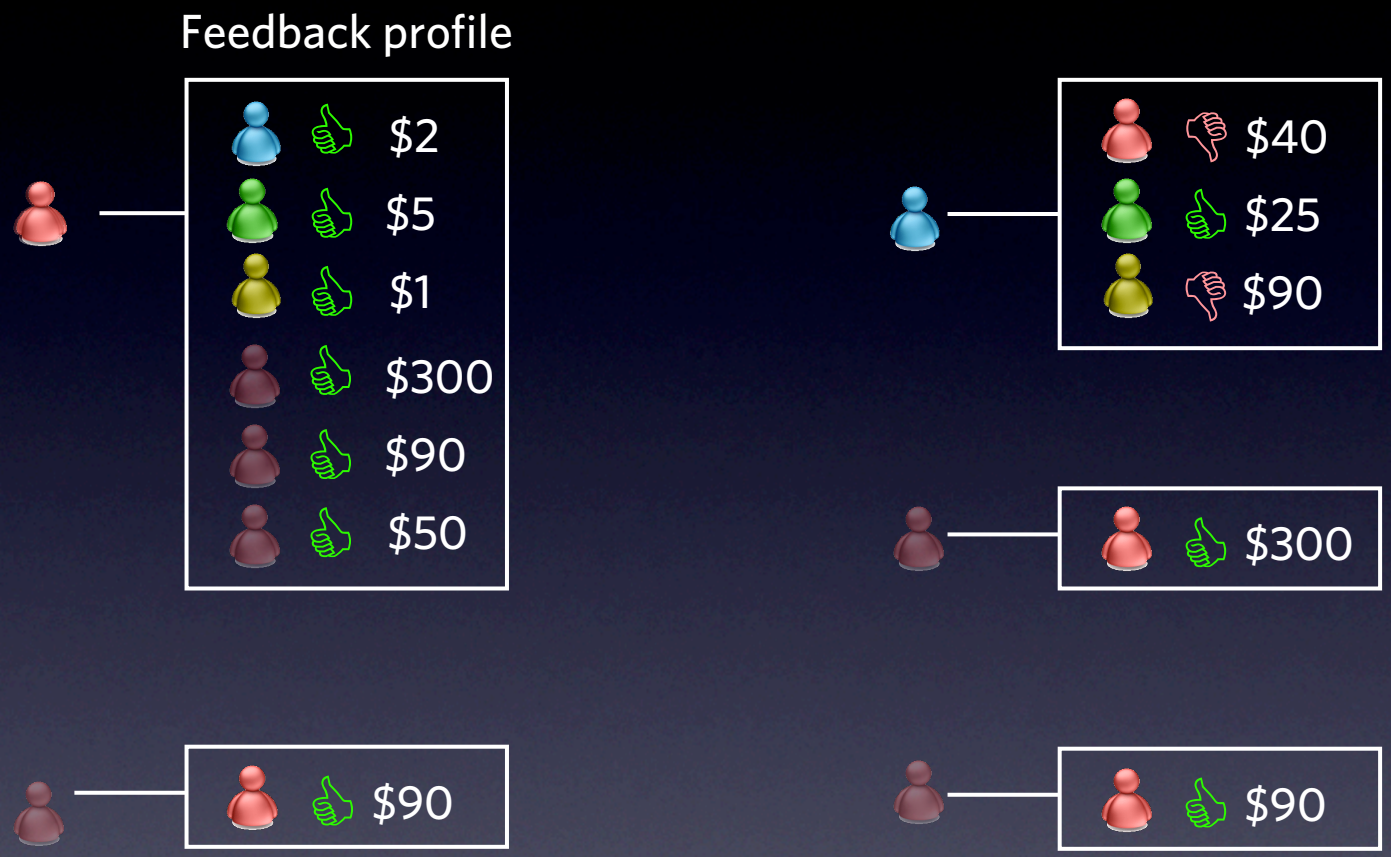
Identities and reputations



Significant monetary losses

- Recent arrest of user who stole $717 κ from 5,000 users
- Used >250 accounts

Identities and reputations



Significant monetary losses

- Recent arrest of user who stole $717 κ from 5,000 users
- Used >250 accounts

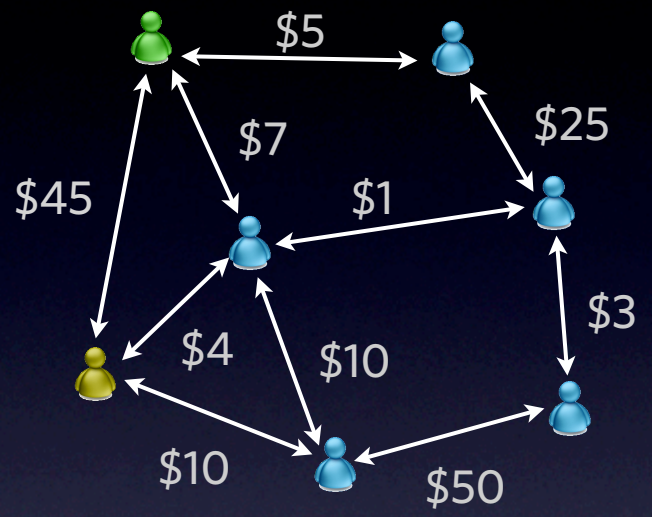
Bazaar: A new approach

New approach to strengthening user reputations
- Leverages an (existing) risk network
- Focuses on protecting buyers from malicious sellers

Works in conjunction with existing marketplace
- Assumes same feedback system as today
- No additional monetary cost
- No strong identities

Insight: Successful transactions represent shared risk
- Buyer and seller more likely to enter into future transactions

Bazaar's risk network



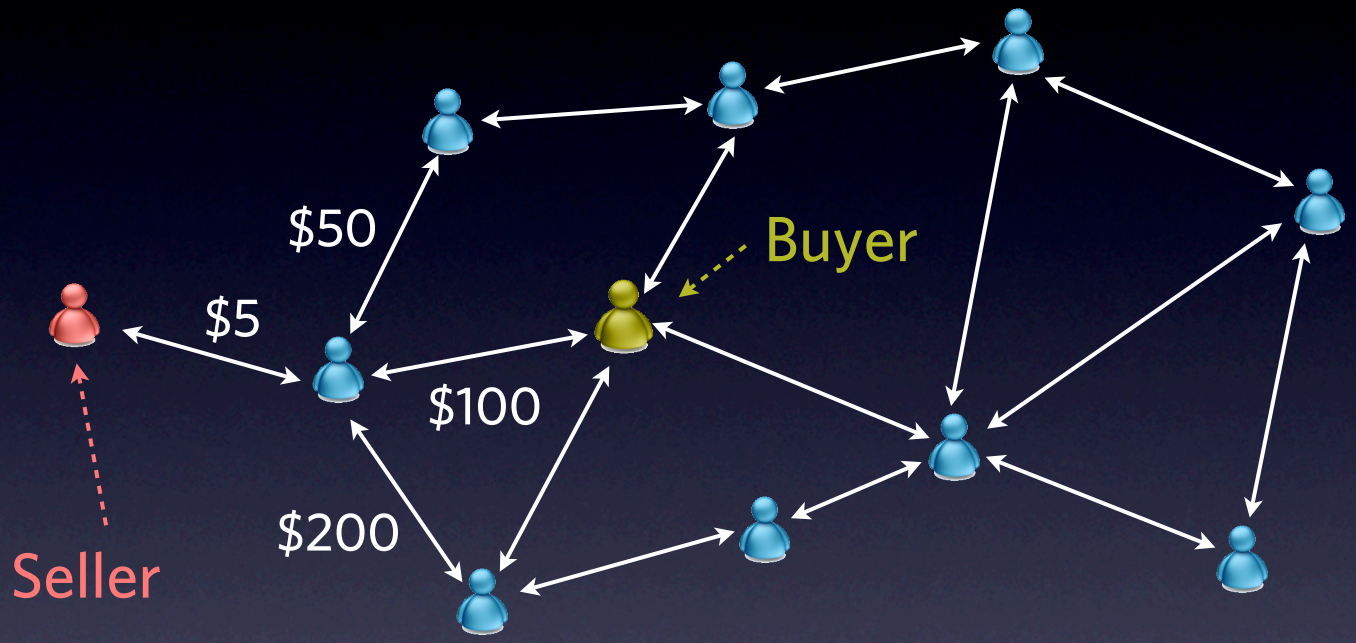
Successful transaction → two identities linked

Weighted by amount of transaction

Risk network automatically generated

Users need not even know about it

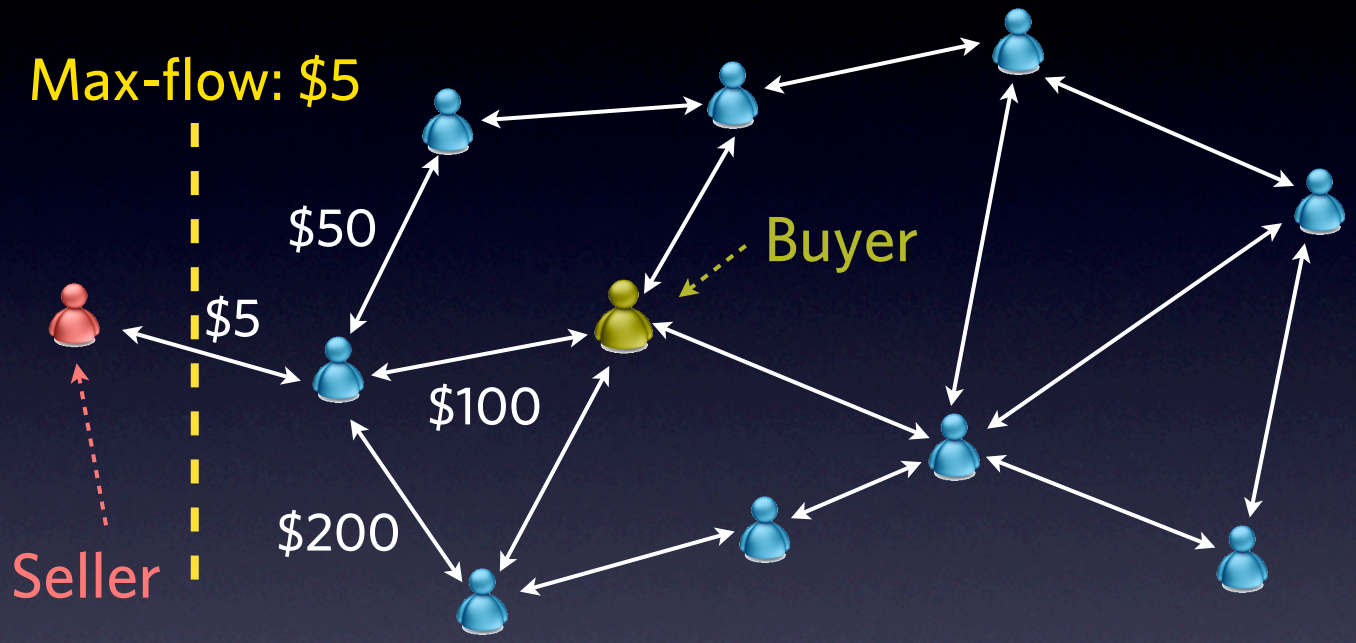
Estimating risk



Bazaar calculates max-flow between buyer and seller

If max-flow lower than potential transaction, flag as fraudulent

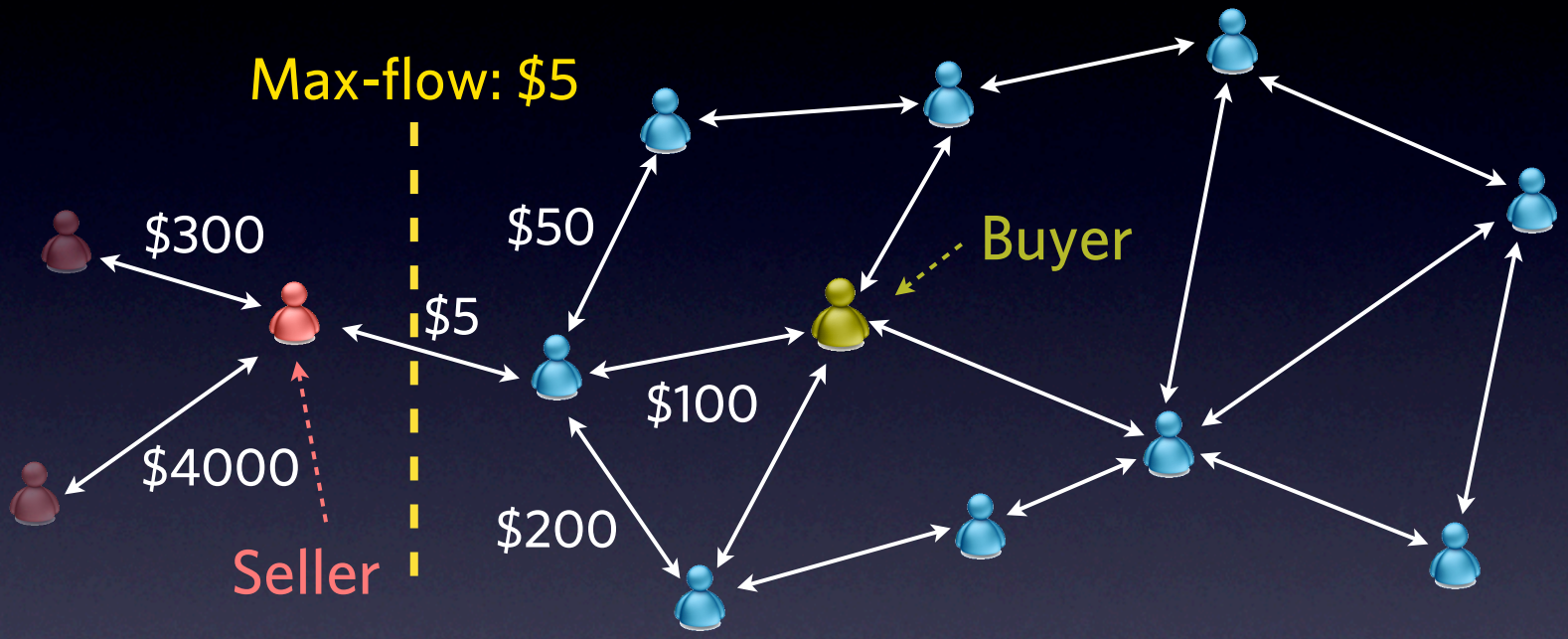
Estimating risk



Bazaar calculates max-flow between buyer and seller

If max-flow lower than potential transaction, flag as fraudulent

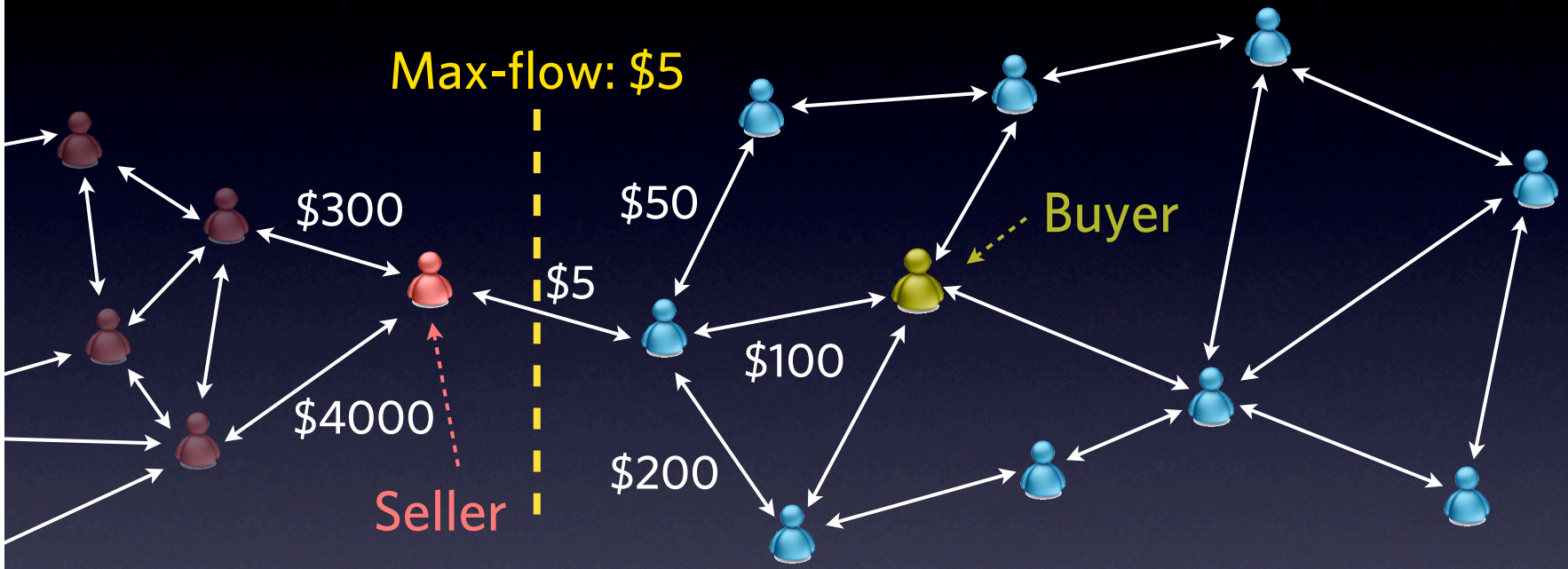
Estimating risk



Bazaar calculates max-flow between buyer and seller

If max-flow lower than potential transaction, flag as fraudulent

Estimating risk



Bazaar calculates max-flow between buyer and seller

If max-flow lower than potential transaction, flag as fraudulent

Summary

Increasing trend of online services with free accounts
- Opens new vector for attack

Focused on reputation manipulation in online marketplaces
- Bazaar: A new approach to strengthening reputations

Evaluated on 10 M auctions from eBay UK
- Would have prevented £164 K of negative feedback
- Only in five categories over 90 days

Currently looking to apply techniques to other domains

Conclusion

Social networks and computer systems increasingly integrated
- New way of organizing information
- Leading to new opportunities, challenges

My group's goal: Leverage social networks in systems design

WebCloud: Addresses challenges with emerging workloads
Bazaar: Addresses challenges with free accounts

Questions?

Work done in collaboration with

Ben Adams *(MPI-I)*, Bobby Bhattacharjee *(University of Maryland)*, Meeyoung Cha *(KAIST)*, Peter Druschel *(MPI-SWS)*, Krishna P. Gummadi *(MPI-SWS)*, Andreas Haeberlen *(University of Pennsylvania)*, Ancsa Hannák *(Northeastern University)*, Jonathan Katz *(University of Maryland)*, Hema Swetha Koppula *(Yahoo Research India)*, Sune Lehmann *(TU Copenhagen)*, Yabing Liu *(Northeastern University)*, Arash Molavi *(Northeastern University)*, Jukka-Pekka Onnela *(Harvard University)*, Ansley Post *(Google)*, J. Niels Rosenquist *(Harvard Medical School)*, Neil Spring *(University of Maryland)*, Ravi Sundaram *(Northeastern University)*, Malveeka Tewari *(University of California, San Diego)*, Bimal Viswanath *(MPI-SWS)*, Liang Zhang *(Northeastern University)*, Fangfei Zhou *(Northeastern University)*