PARADIGM SHIFT:

DATA TOGETHER

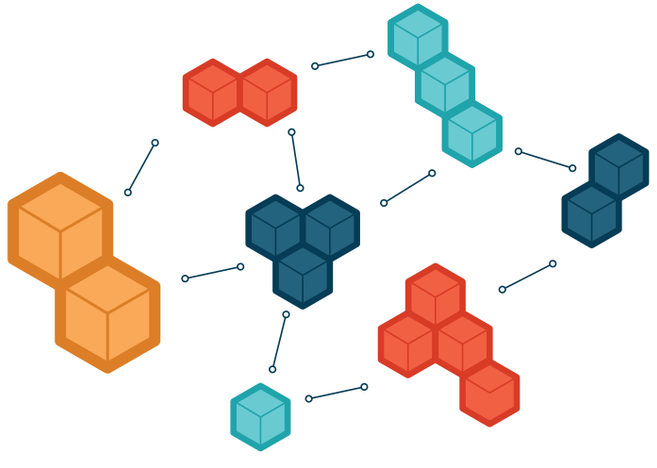
COMMUNITIES & INSTITUTIONS

USING DECENTRALIZED TECHNOLOGIES TO MAKE A BETTER WEB

MATT ZUMWALT
PROTOCOL LABS

PASIG AUTUMN 2017

OXFORD UNIVERSITY
12 SEPTEMBER 2017



Protocol Labs

DECENTRALIZED WEB

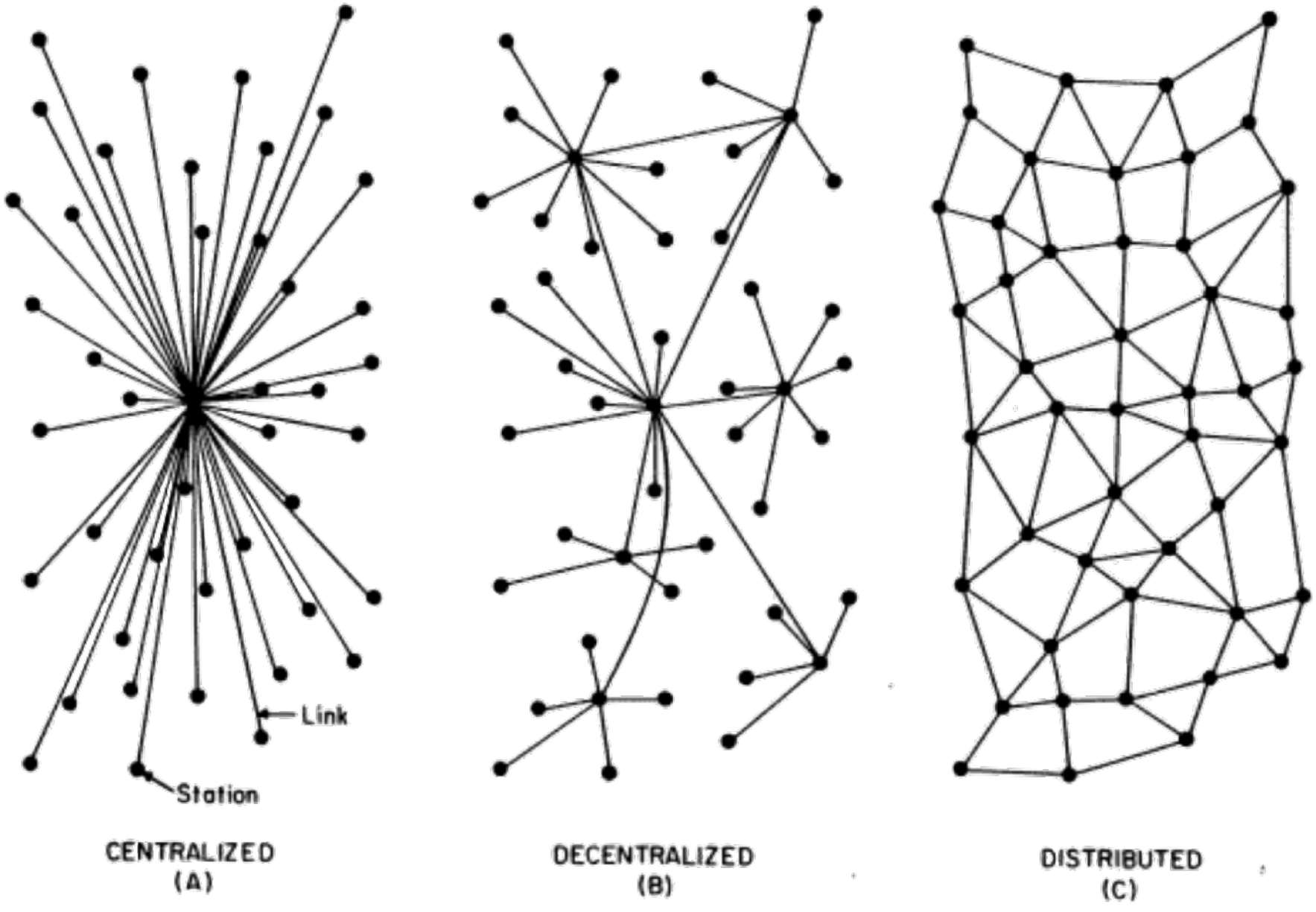


FIG. 1 — Centralized, Decentralized and Distributed Networks

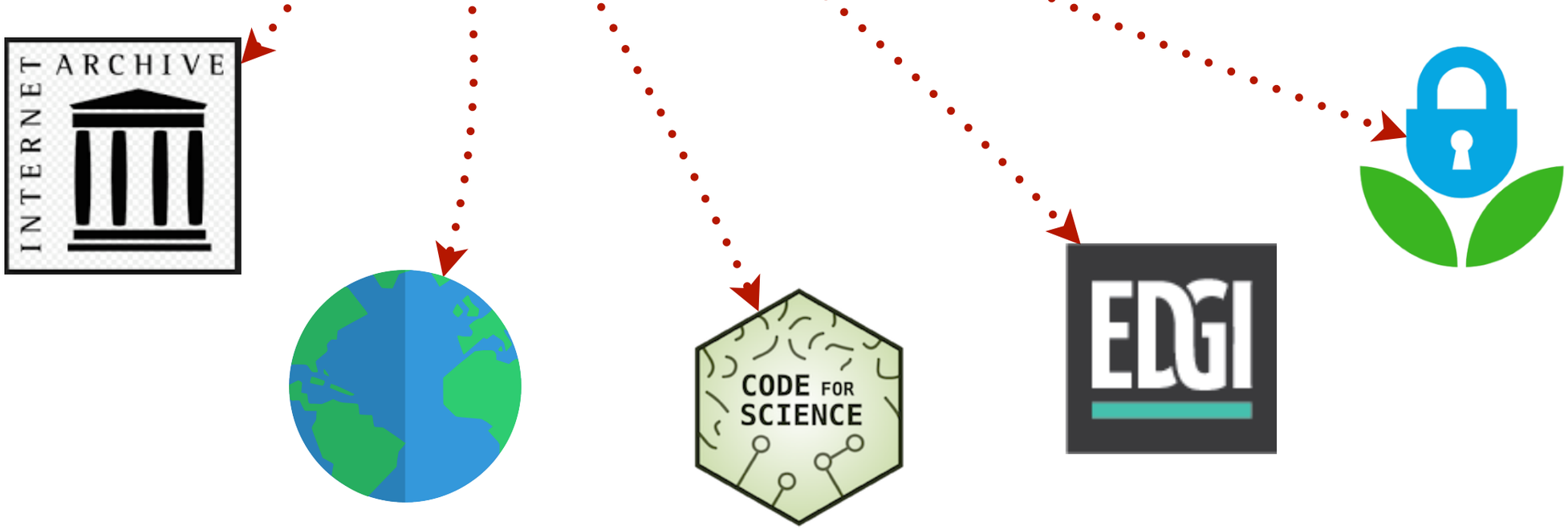
Source: On Distributed Communications Networks, Paul Baran, 1962

The internet has been stolen from you. Take it back, nonviolently.

https://medium.com/@flyingzumwalt/the-internet-has-been-stolen-from-you-take-it-back-nonviolently-248f8d445b87

re-contextualizing

DATA RESCUE



re-contextualizing

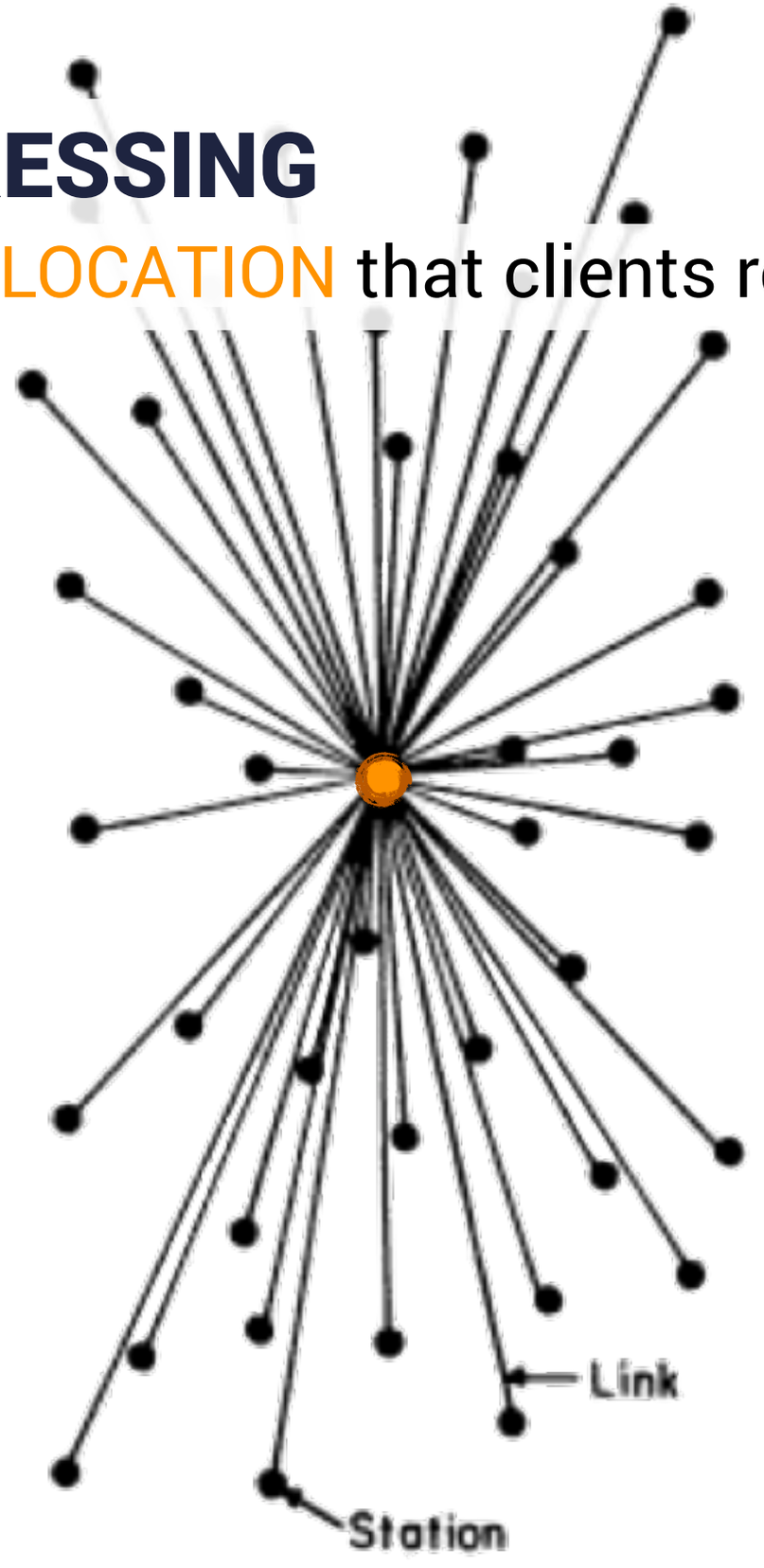
DATA RESCUE

THOUSANDS OF PEOPLE SPENDING DAYS, WEEKS OR MONTHS OF THEIR PERSONAL TIME TRYING TO PRESERVE OTHER PEOPLE'S DATA

confounded by

LOCATION-ADDRESSING

competing to be THE LOCATION that clients rely on

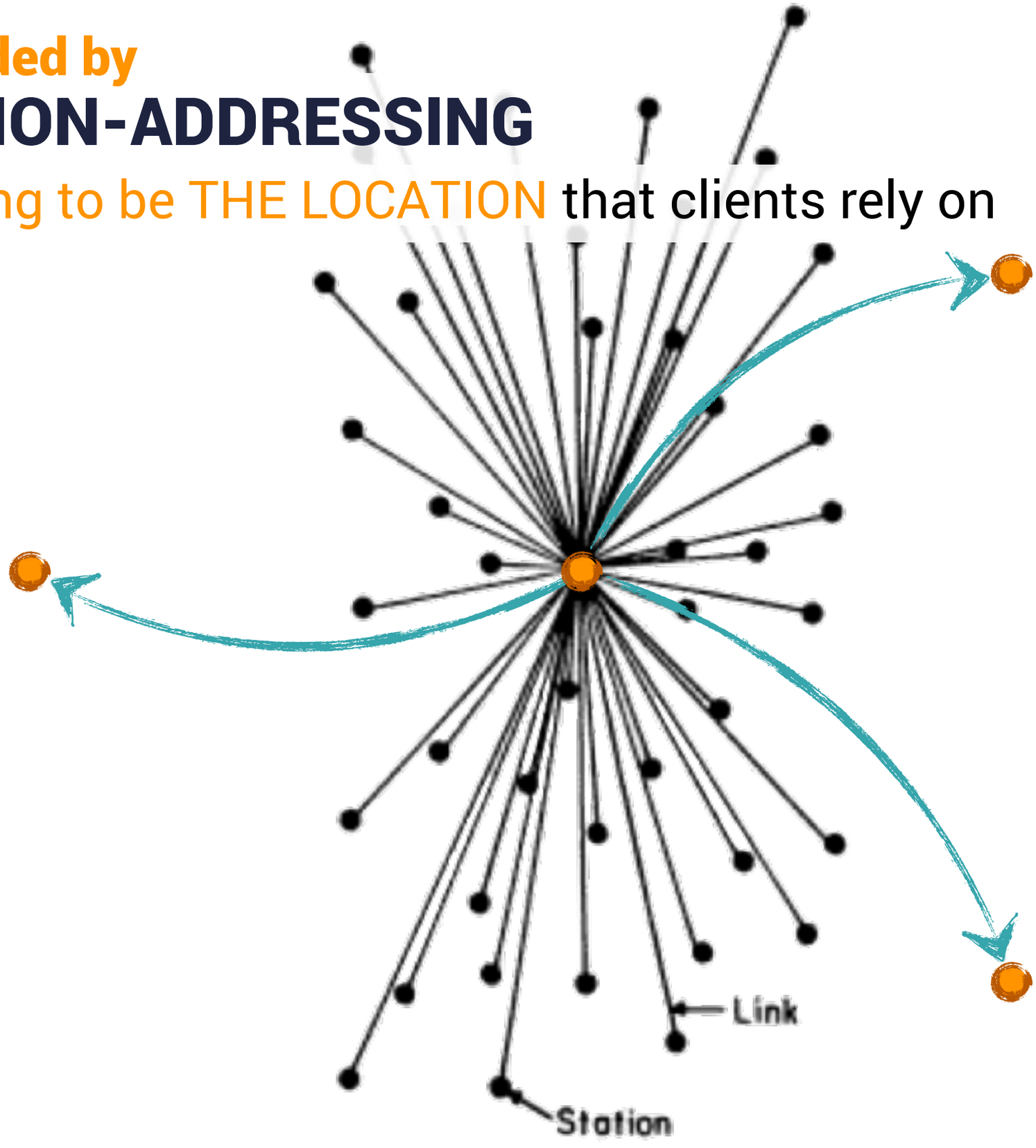


CENTRALIZED
(A)

confounded by

LOCATION-ADDRESSING

competing to be THE LOCATION that clients rely on

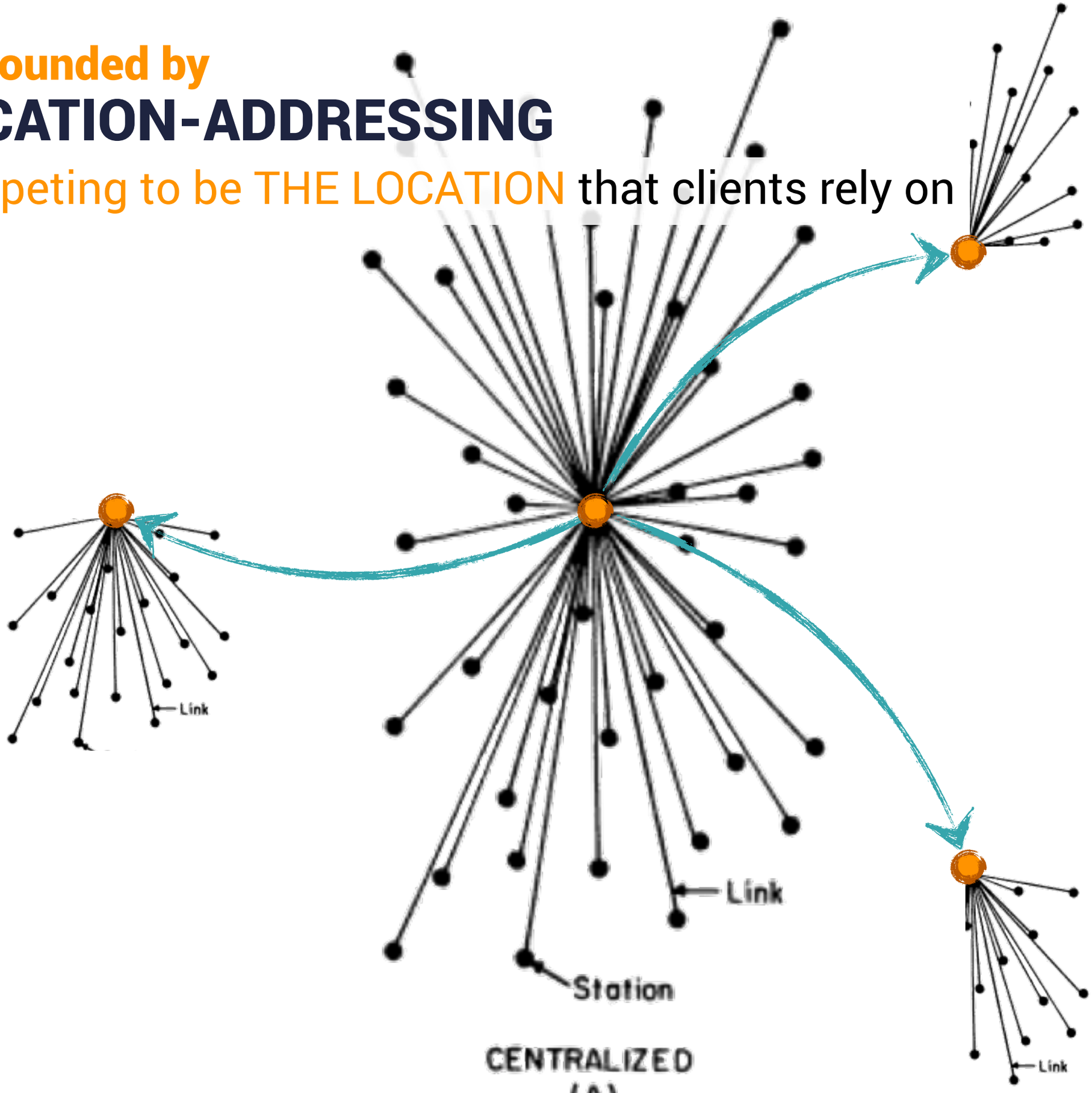


CENTRALIZED
(A)

confounded by

LOCATION-ADDRESSING

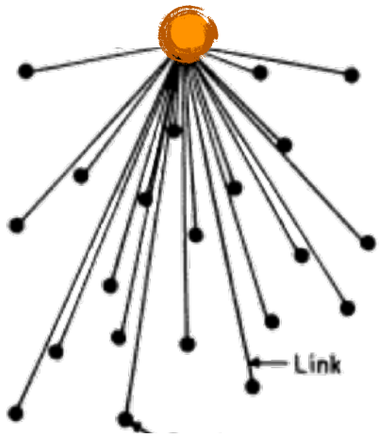
competing to be THE LOCATION that clients rely on

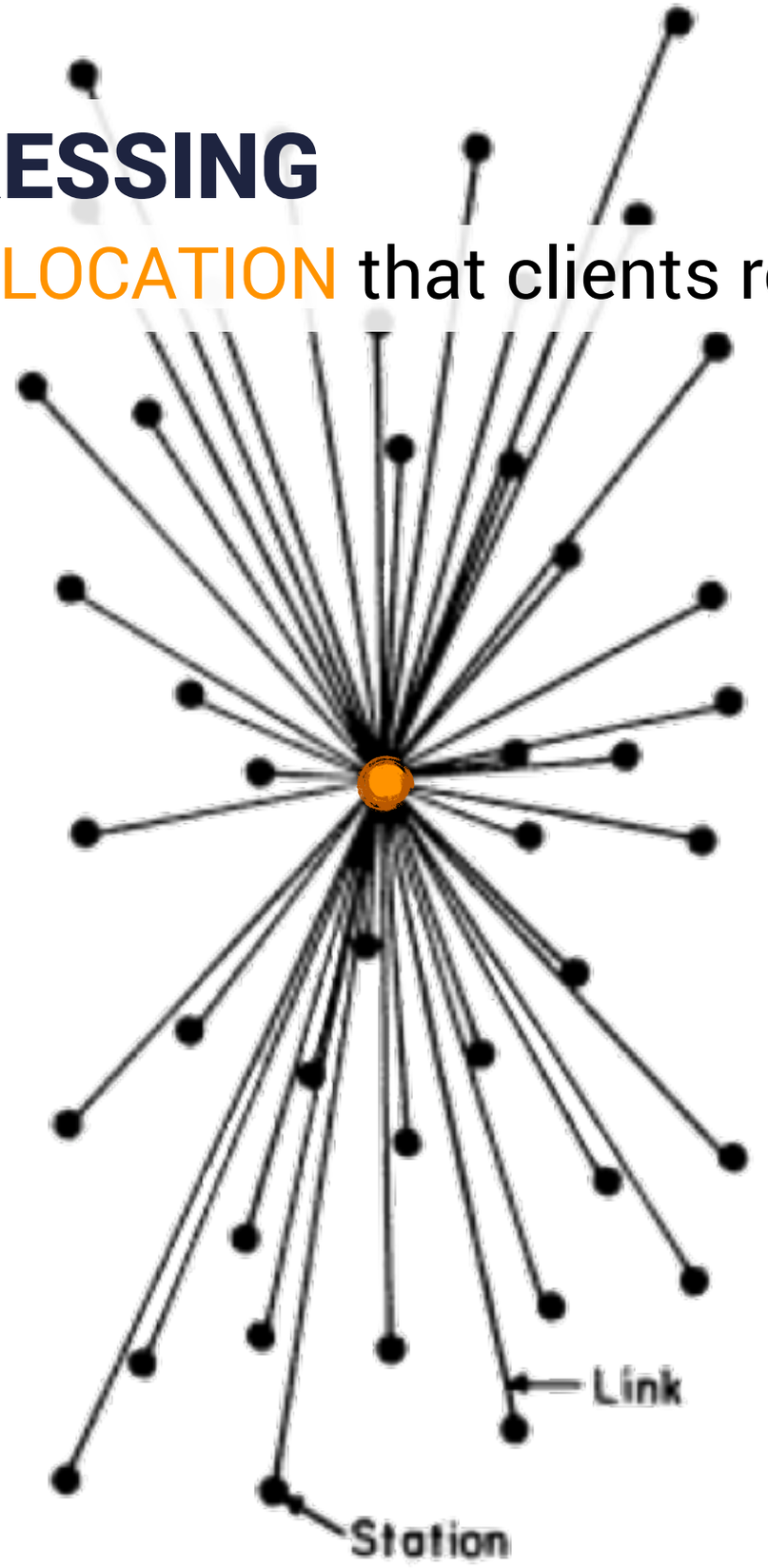


confounded by

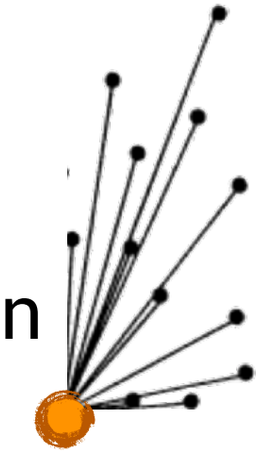
LOCATION-ADDRESSING

competing to be THE LOCATION that clients rely on





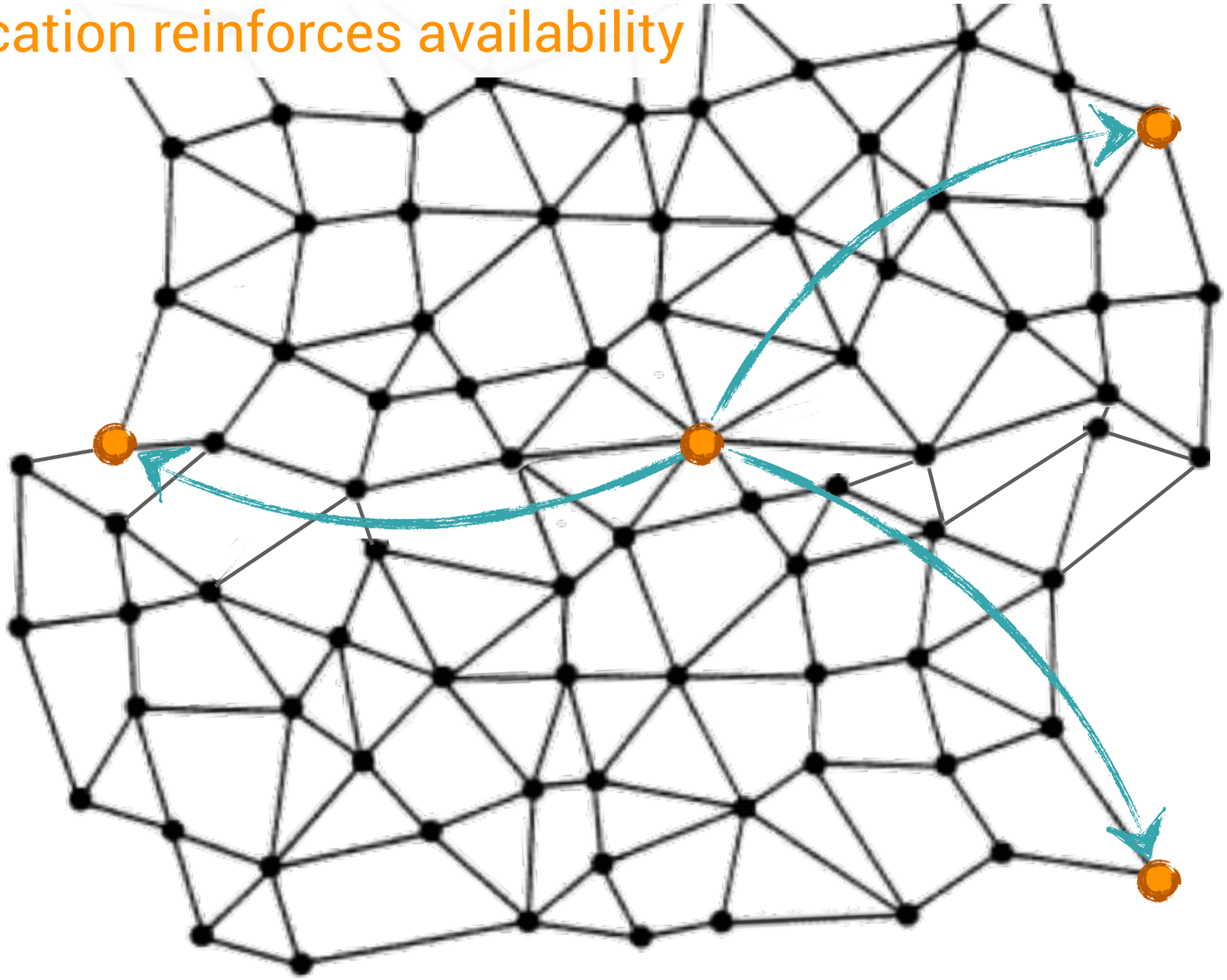
CENTRALIZED
(A)



a better web

PEER TO PEER AND CONTENT-ADDRESSED

replication reinforces availability

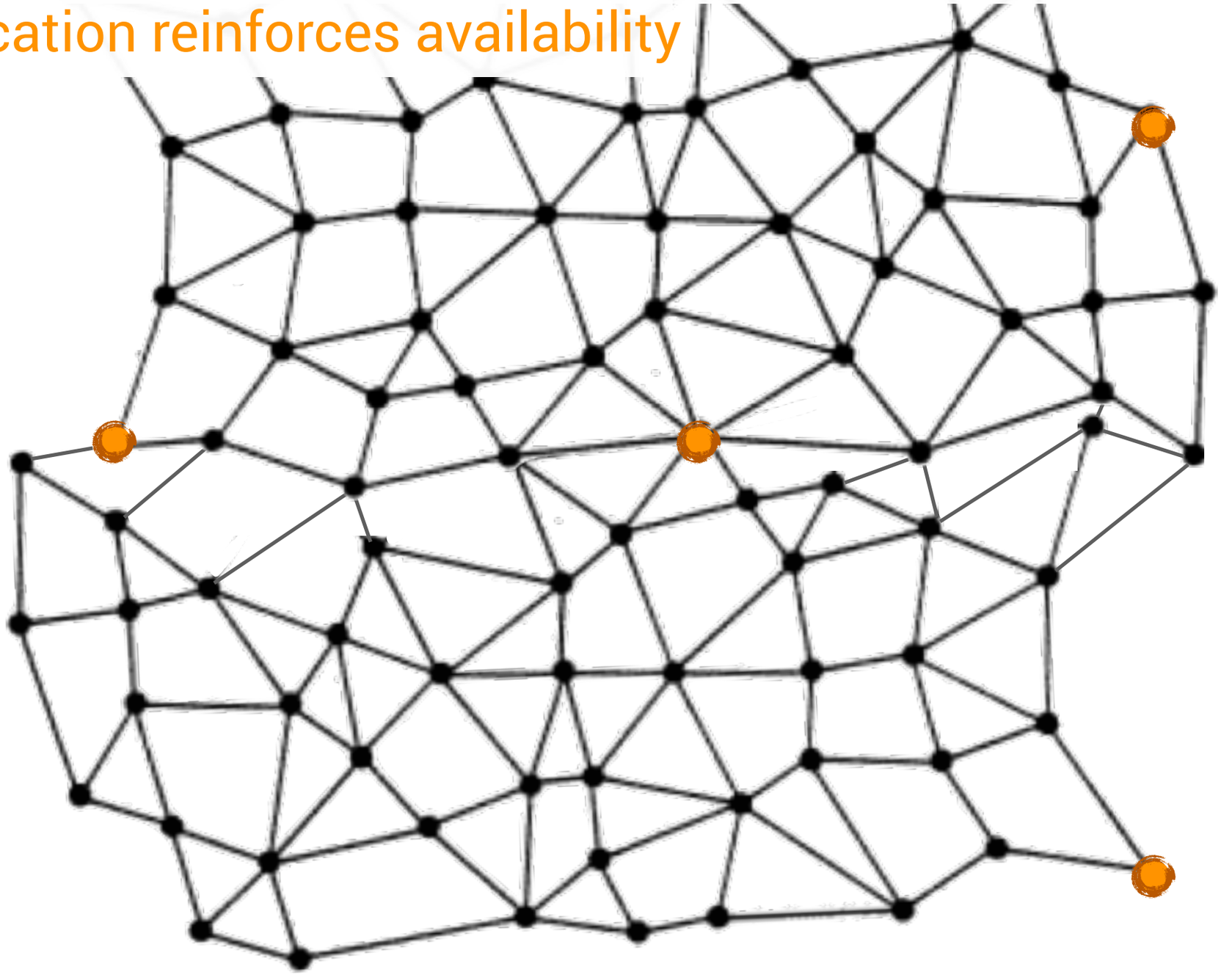


replicating data is easy

a better web

PEER TO PEER AND CONTENT-ADDRESSED

replication reinforces availability

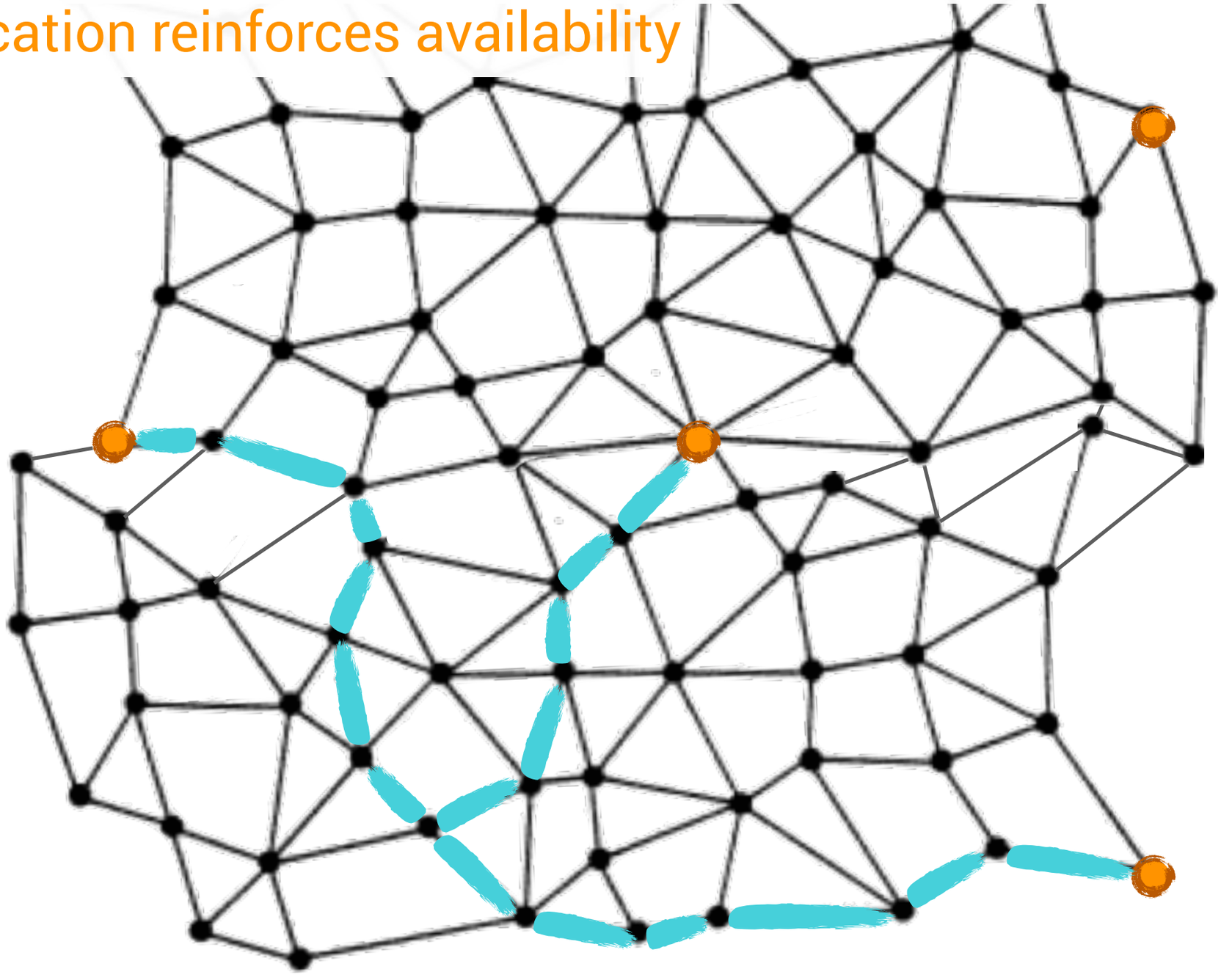


validating data is easy

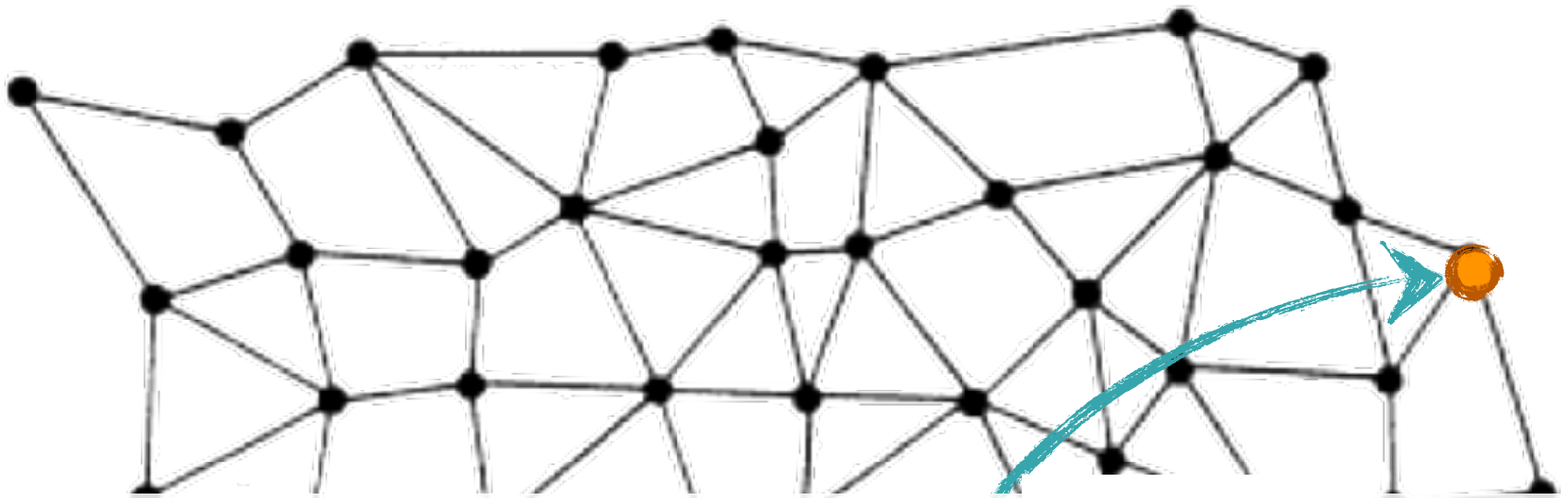
a better web

PEER TO PEER AND CONTENT-ADDRESSED

replication reinforces availability



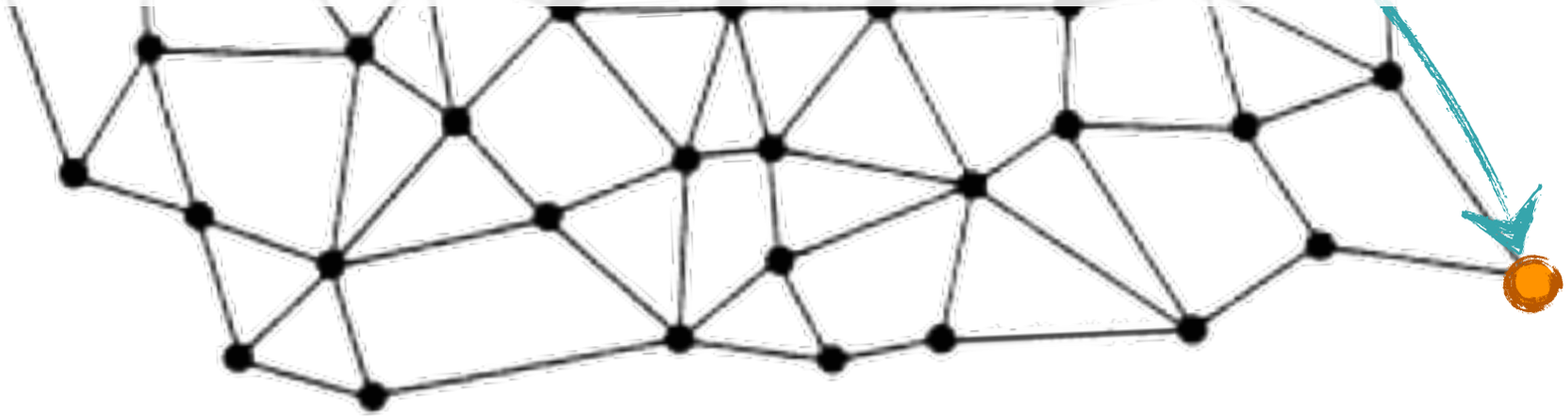
replicating data makes more paths for nodes to access content



MATRIX OF RISK + VALUE

can’t be built by any one organization or one community

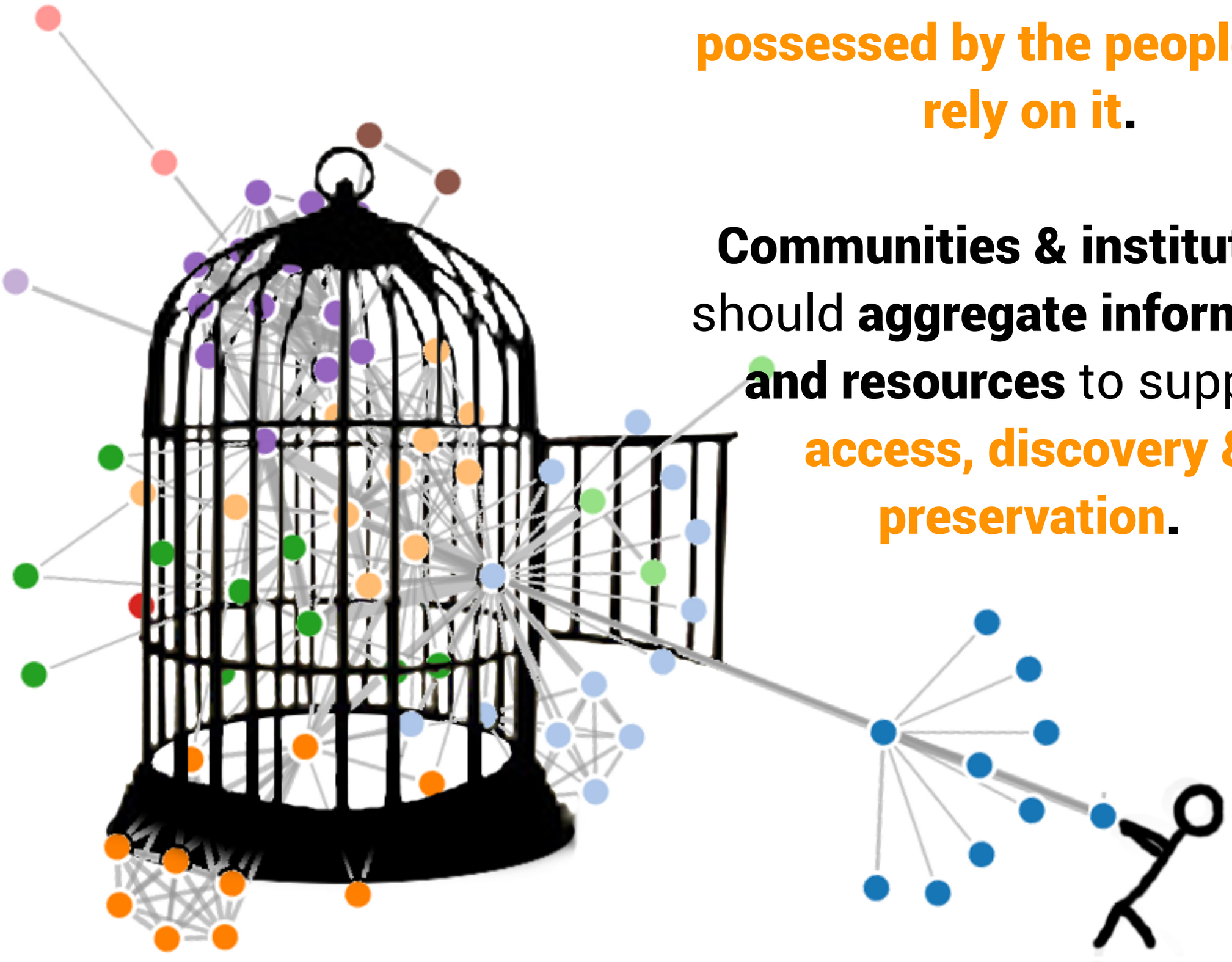
and shouldn’t be.



access, discovery and preservation are participatory

Information should be **possessed by the people who rely on it**.

Communities & institutions should **aggregate information and resources** to support **access, discovery & preservation**.



https://medium.com/@flyingzumwalt/the-internet-has-been-stolen-from-you-take-it-back-nonviolently-248f8d445b87

SHARING POSSESSION OF INFORMATION RESOURCES

LOCATION-ADDRESSING

same info in 100 locations == 100 information resources

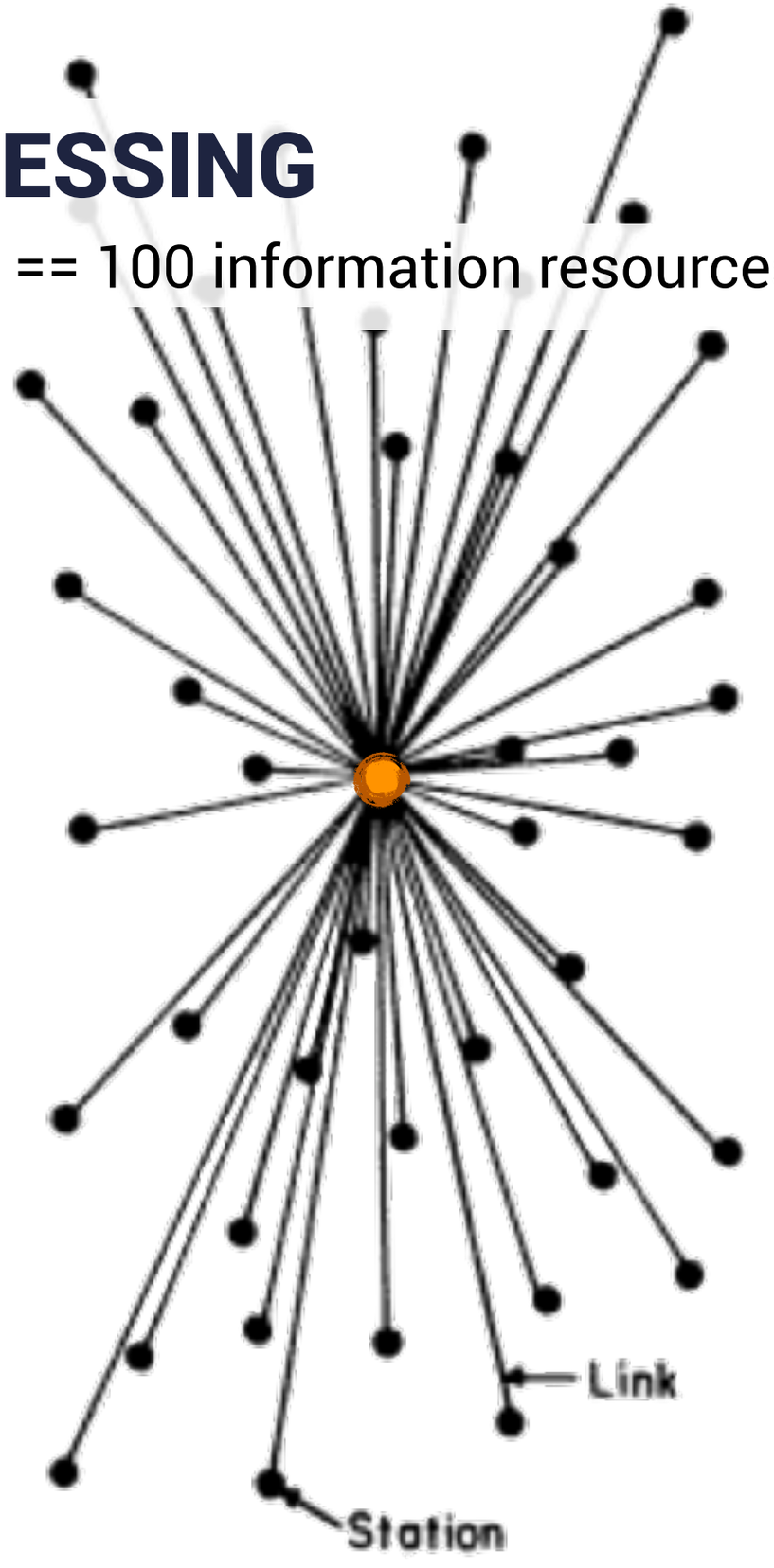
SHARING POSSESSION OF INFORMATION RESOURCES

CONTENT-ADDRESSING

same info in 100 locations == 1 information resource, possessed by many

LOCATION-ADDRESSING

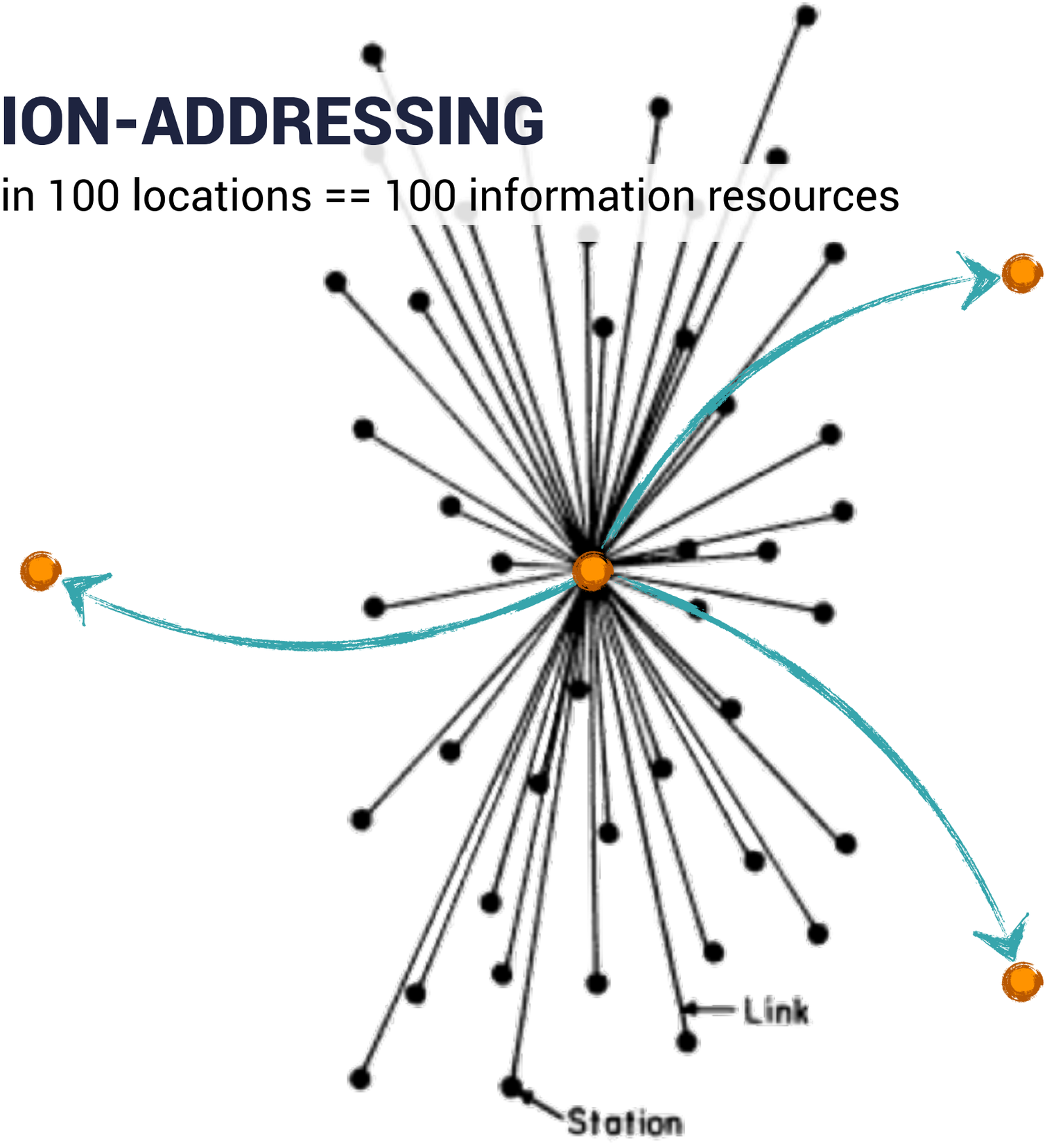
same info in 100 locations == 100 information resources



CENTRALIZED
(A)

LOCATION-ADDRESSING

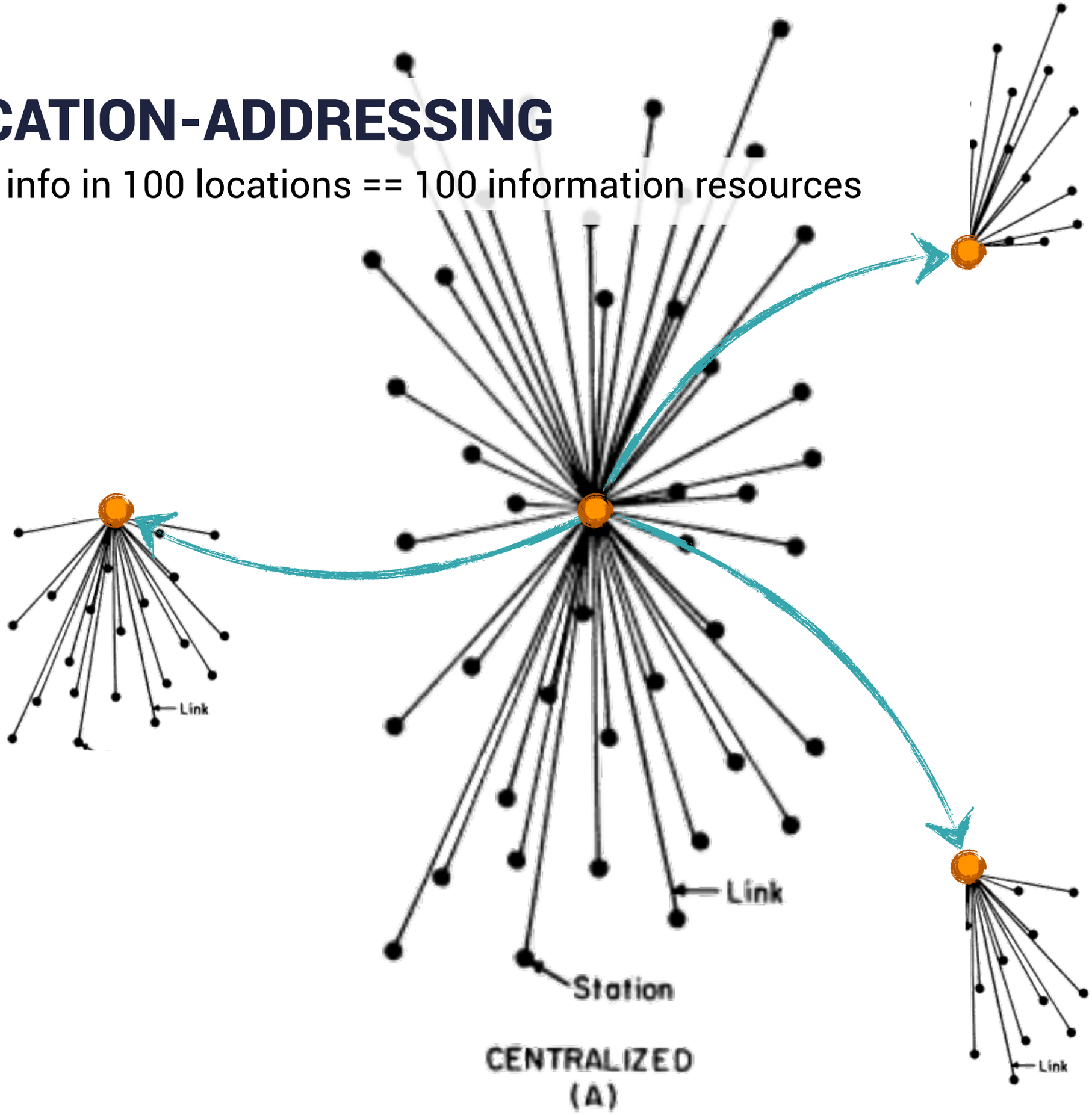
same info in 100 locations == 100 information resources



CENTRALIZED
(A)

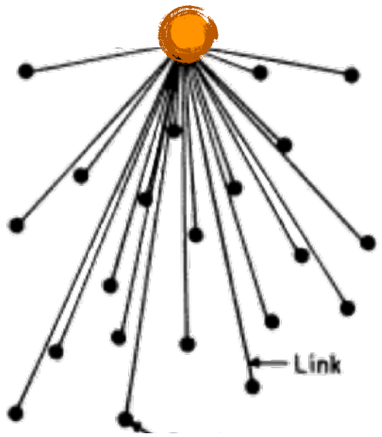
LOCATION-ADDRESSING

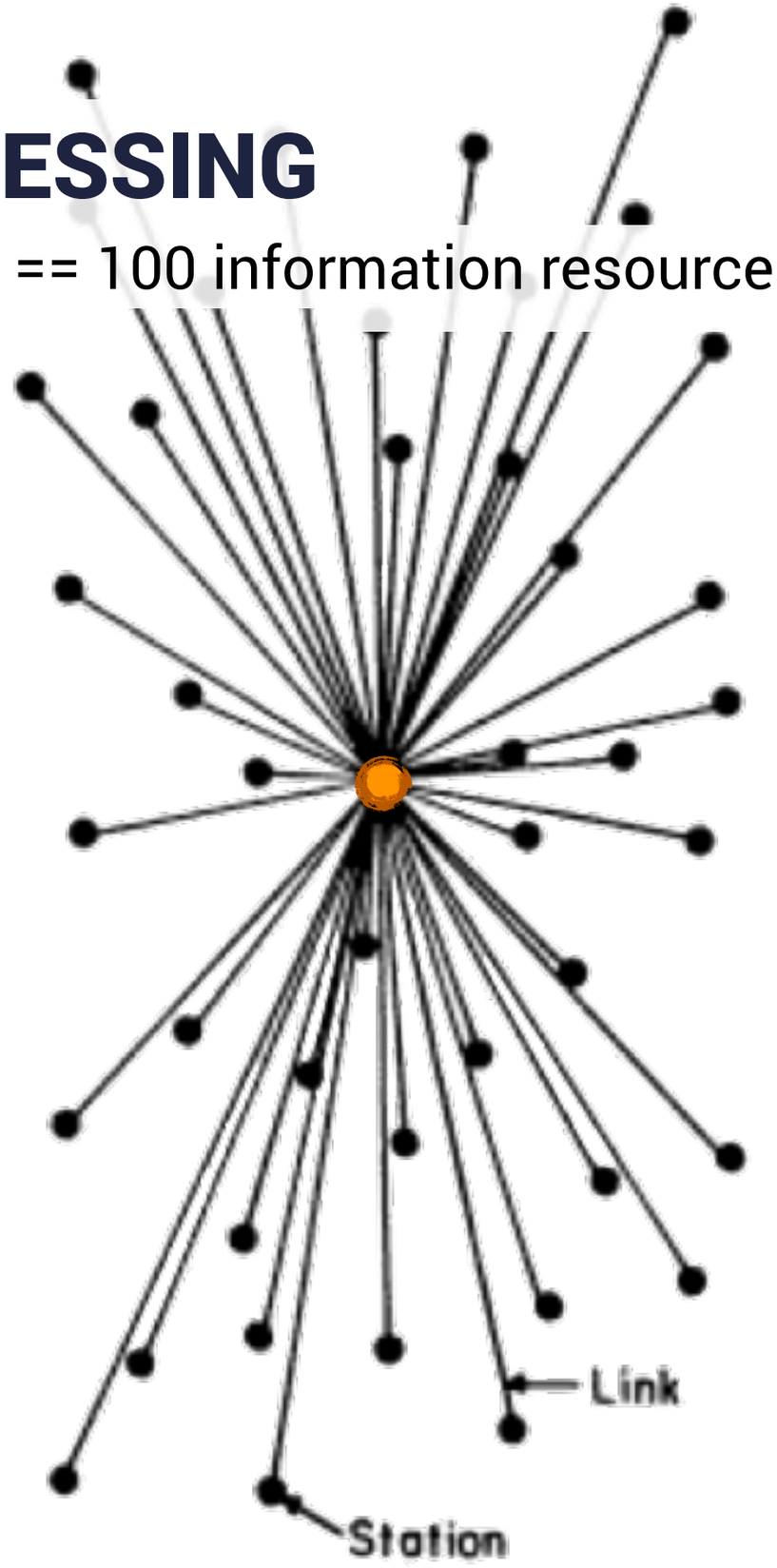
same info in 100 locations == 100 information resources



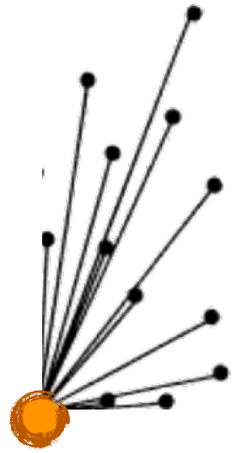
LOCATION-ADDRESSING

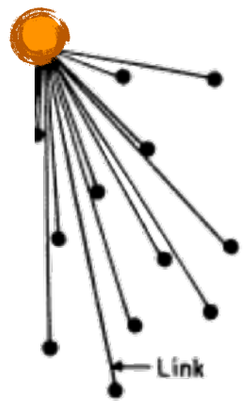
same info in 100 locations == 100 information resources





CENTRALIZED
(A)





CORE CONCEPT

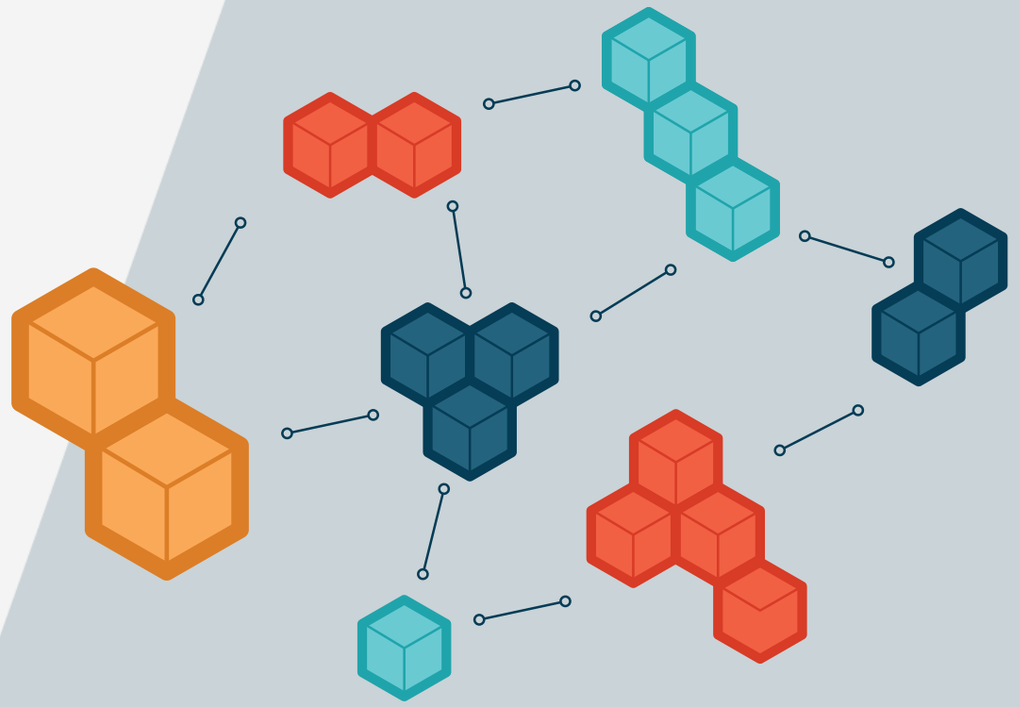
CONTENT ADDRESSING

Benefit: If *anyone on the network* has a copy of the *content you will be able to find and retrieve it*

CORE CONCEPT

CONTENT ADDRESSING

Key idea: It doesn't matter where the content is stored. What matters is being able to know that you're getting exactly the content you requested*

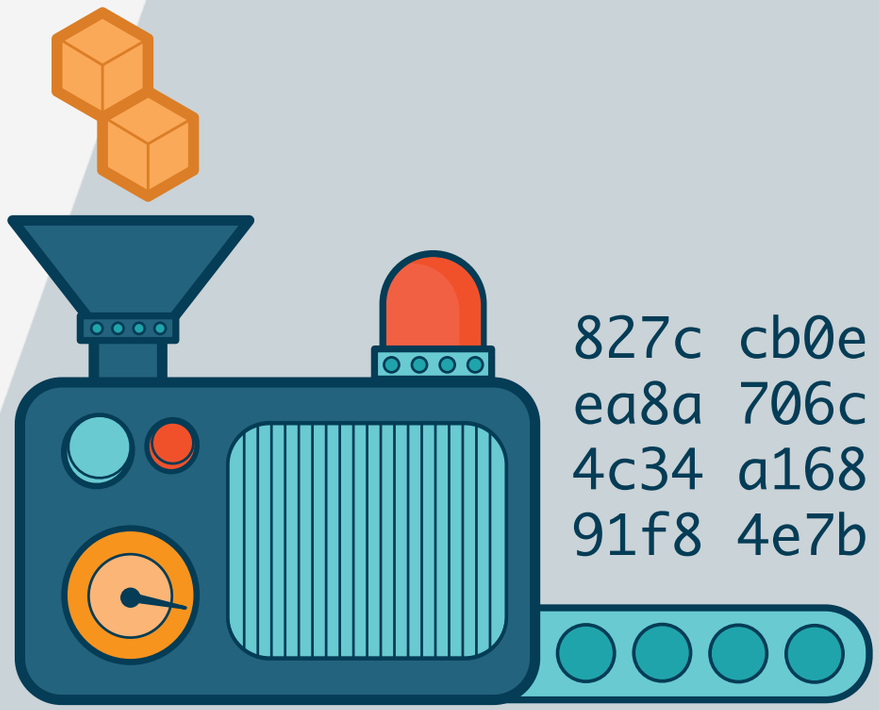


*also important to consider who you got the link/address from.

CORE CONCEPT

CONTENT ADDRESSING

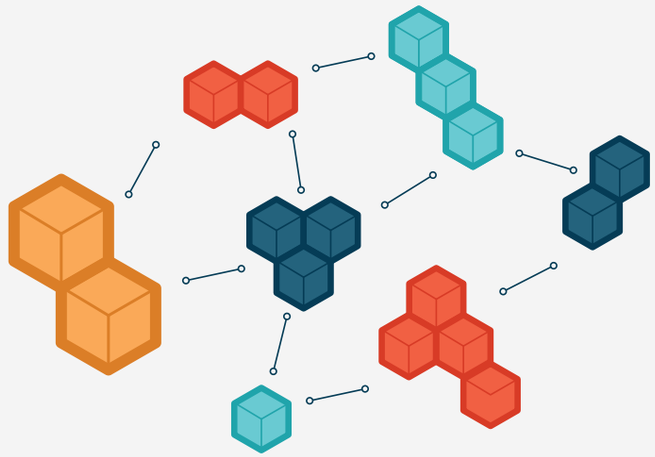
How: identify content by its cryptographic hash



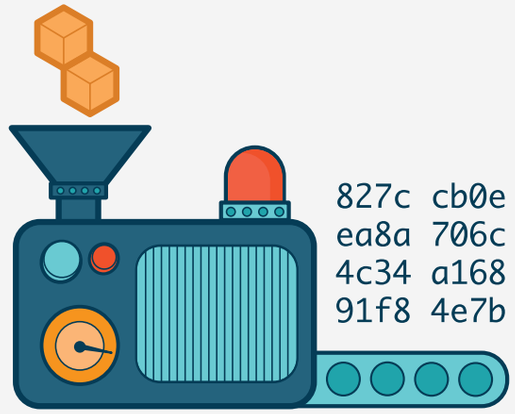
CORE CONCEPT

CONTENT ADDRESSING

Benefit: If *anyone on the network* has a copy of the *content you will be able to find and retrieve it*



Key idea: It doesn't matter where the content is stored. What matters is being able to know that you're getting exactly the content you requested*



How: Identify content by its cryptographic hash

*also important to consider who you got the link/address from.

P2P ACCESSION OF DATA

This file is important.

“Can we pin a copy at the library?”

We’ve got you covered. **Just tell me the hash**. We will accession all of it.

LIBRARIAN

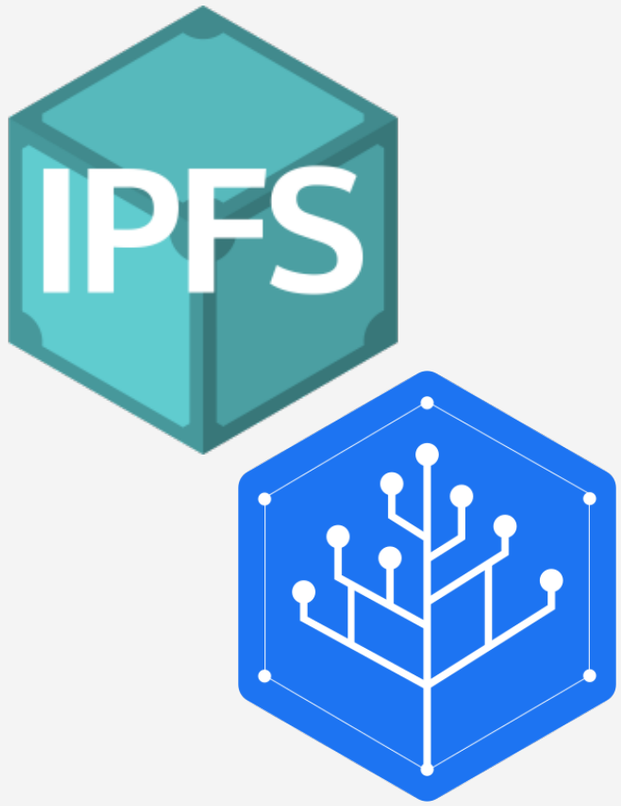
ACCESSION FEEDS OF DATA

“Please pin all the versions of my dataset. The version history is in the feed at dweb:/ipns/jack/dataset1.

I’ll keep updating it when there are new versions of my data.”

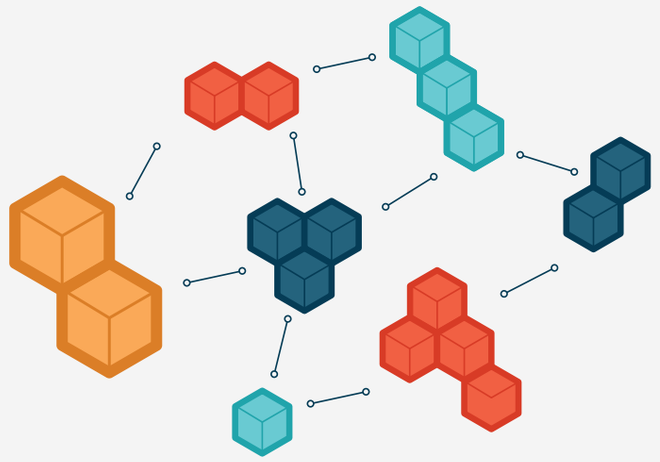
“No problem. We’ve pinned all the versions in that feed and will grab any new versions when they show up.”

LIBRARIAN



THE GAME CHANGER:

HASH-LINKED DATA STRUCTURES



SOME TOOLS THAT USE

HASH-LINKED DATA STRUCTURES

A BENEFIT OF HASH-LINKING:

DECOUPLE CONTENT FROM LOCATION

DECOUPLE *WHERE THE CONTENT IS STORED* FROM THE IDENTITY OF THAT CONTENT, SO THAT **DATA CAN EXIST IN MANY PLACES AND PASS THROUGH MANY HANDS** WITHOUT LOSING INTEGRITY

A BENEFIT OF HASH-LINKING:

CRYPTOGRAPHIC INTEGRITY CHECKING

WHEN RESOLVING A LINK **YOU CAN USE THE LINK VALUE (A HASH) TO VALIDATE THE RESULT**. THIS ALLOWS WIDE, SECURE, EXCHANGES OF DATA (E.G. GIT OR BITTORRENT)

A BENEFIT OF HASH-LINKING:

IMMUTABLE DATA STRUCTURES

DATA STRUCTURES WITH HASH LINKS **CANNOT MUTATE**. THIS IS USEFUL FOR VERSIONING, FOR REPRESENTING DISTRIBUTED MUTABLE STATE, AND FOR LONG TERM ARCHIVING.

PUTTING IT TOGETHER

CONTENT-ADDRESSED PROTOCOL

adding via command line:

```
$ ipfs add -r path-to-data-youre-adding
added  QmXoypizj...
```

retrieving via command line:

```
$ ipfs get QmXoypizj.../a-path-inside-dataset
```

http gateway:

https://your-ipfs-node/ipfs/QmXoypizj.../a-path

PUTTING IT TOGETHER

CONTENT-ADDRESSED PROTOCOL

Real Hashes!

adding via command line:

```
$ ipfs add -r ./wikipedia-en
added QmXoypizjW3WknFiJnKLwHCnL72vedxjQkDDP1mXWo6uco
```

retrieving via command line:

```
$ ipfs get QmXoypizjW3WknFiJnKLwHCnL72vedxjQkDDP1mXWo6uco/wiki/Peer-to-peer.html
```

retrieving via http gateway:

https://gateway.ipfs.io/ipfs/
QmXoypizjW3WknFiJnKLwHCnL72vedxjQkDDP1mXWo6uco
/wiki/Peer-to-peer.html

PUTTING IT TOGETHER

PINNING DATA ON A NODE

$ ipfs pin add QmXoypizj...

Pinning a hash tells an IPFS node to keep the corresponding content. Otherwise the content will be deleted when the node runs its garbage collection routine.

pin sets as collections

CONTENT-ADDRESSED P2P DATA COLLECTIONS

BENEFICIAL IMPACTS ON ACCESSIONING, ACCESS, DISCOVERY, AND PRESERVATION

pin sets as collections

IMPACT ON PRESERVATION

replication is easy and transparent

downloads are replicas

participatory preservation

integrity checking is automatic

easier format migrations

content-addressed versioning

pin sets as collections

IMPACT ON ACCESSIONING

anticipates next generation of tools

submission without upload

easier to provide context

submit versioned series of hashes

pin sets as collections

IMPACT ON ACCESS

replication reinforces availability

availability scales with demand

automatic integrity checking

specificity about versions

pin sets as collections

IMPACT ON DISCOVERY

enables “collections as data”

share versioned metadata via hash

supports machine analysis

forking collections !?!

powerful means of deduplication

https://datatogether.org

PARADIGM SHIFT:

DATA TOGETHER

COMMUNITIES & INSTITUTIONS

USING DECENTRALIZED TECHNOLOGIES TO MAKE A BETTER WEB

MATT ZUMWALT
PROTOCOL LABS

NDSR SYMPOSIUM

WORLD BANK
17 AUGUST 2017