Overlay (and P2P) Networks

Part II

- Recap (Amazon Dynamo)
- Error and Attack Tolerance of Complex Networks

Samu Varjonen

Ashwin Rao

Amazon Dynamo

Recap of ACID, BASE

- **A**tomicity, **C**onsistency, **I**solation, **D**urability
- CAP Principle
  - **C**: Strong Consistency (single-copy ACID consistency)
  - **A**: High Availability (available at all times)
  - **P**: Partition Resilience (survive partition between replicas)

PICK ANY TWO

  - Once a transaction has been committed its results, the system must guarantee the results survive subsequent malfunctions
- **B**asically **A**vailable, **S**oft state, **E**ventually consistent

Requirements from Dynamo

G. DeCandia et al. **“Dynamo: Amazon’s Highly Available Key-value Store,”** In SOSP 2007.

Requirements from Dynamo

- Key-value store
  - shopping carts, seller lists, preferences, product catalog

G. DeCandia et al. **"Dynamo: Amazon's Highly Available Key-value Store,"** In SOSP 2007.

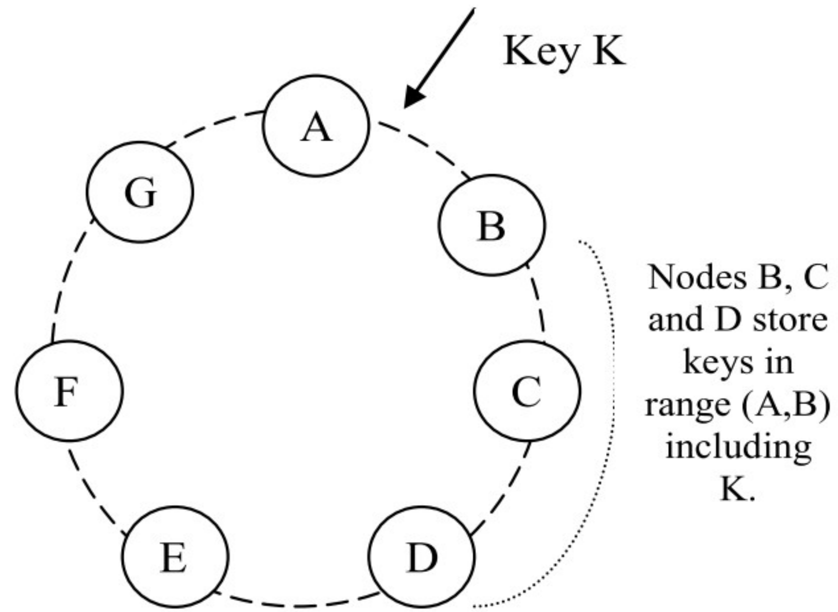
Requirements from Dynamo

- Key-value store
  - shopping carts, seller lists, preferences, product catalog
- System built using off-the-shelf hardware.
- Platform must scale to support continuous growth
- Address tradeoff of high-availability, guaranteed performance, cost-effectiveness, and performance

G. DeCandia et al. **"Dynamo: Amazon's Highly Available Key-value Store,"** In SOSP 2007.

Partitioning and Replication in Dynamo

- Consistent Hashing DHT
  - Virtual nodes in DHT
  - Each physical node added as multiple virtual nodes
- Each data-item replicated in N nodes
  - Each virtual node responsible for the region between it and its Nth predecessor
  - Preference List: list of nodes (in (multiple datacenters) storing a key



G. DeCandia et al. **"Dynamo: Amazon's Highly Available Key-value Store,"** In SOSP 2007.

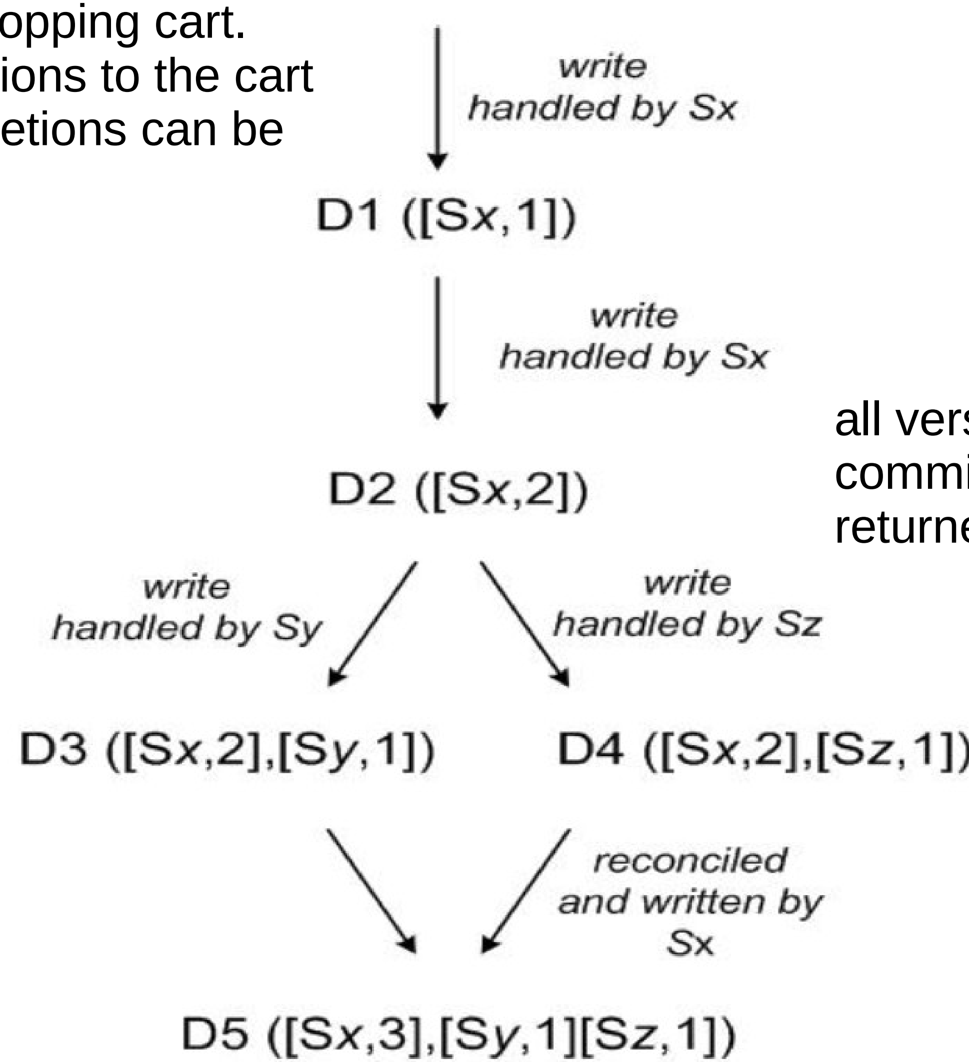
API

- **get** (key)
  - may return many versions of the same object
- **put**(key, ***context***, object)
  - Context: encodes system metadata and includes information such as the version of the object
  - may return to its caller before the update has been applied at all the replicas
  - An object may have different version sub-histories
- Vector clock based versioning
  - One vector clock associated with every version of objects

Data Versioning

Objects versions: D1, D2, D3, ...

Assume object is shopping cart.
Requirements: additions to the cart don’t get lost but deletions can be lost



all versions of the object committed to the system are returned when read

G. DeCandia et al. **“Dynamo: Amazon’s Highly Available Key-value Store,”** In SOSP 2007.

Sloppy Quorum

Sloppy Quorum

- Read + Write involves N nodes from the preference list
  - R: minimum number of nodes for Read
  - W: minimum number of nodes for Write

Sloppy Quorum

- Read + Write involves N nodes from the preference list
  - R: minimum number of nodes for Read
  - W: minimum number of nodes for Write
- R + W > N
  - R = W = 5 → high consistency but system is vulnerable to network partitions
  - R = W = 1 → weak consistency with failure
  - Typical values of (N, R, W) = (3,2,2) → balance between performance and consistency

Read and Write Operations

Read and Write Operations

- Coordinator
  - Node responsible for read/writes
  - First node in the preference list

Read and Write Operations

- Coordinator
  - Node responsible for read/writes
  - First node in the preference list
- Write Operation

Read and Write Operations

- Coordinator
  - Node responsible for read/writes
  - First node in the preference list
- Write Operation
  - New vector clock from coordinator
  - Write locally and forward to N-1 nodes, if W-1 nodes respond then write was successful

Read and Write Operations

- Coordinator
  - Node responsible for read/writes
  - First node in the preference list
- Write Operation
  - New vector clock from coordinator
  - Write locally and forward to N-1 nodes, if W-1 nodes respond then write was successful
- Read Operation

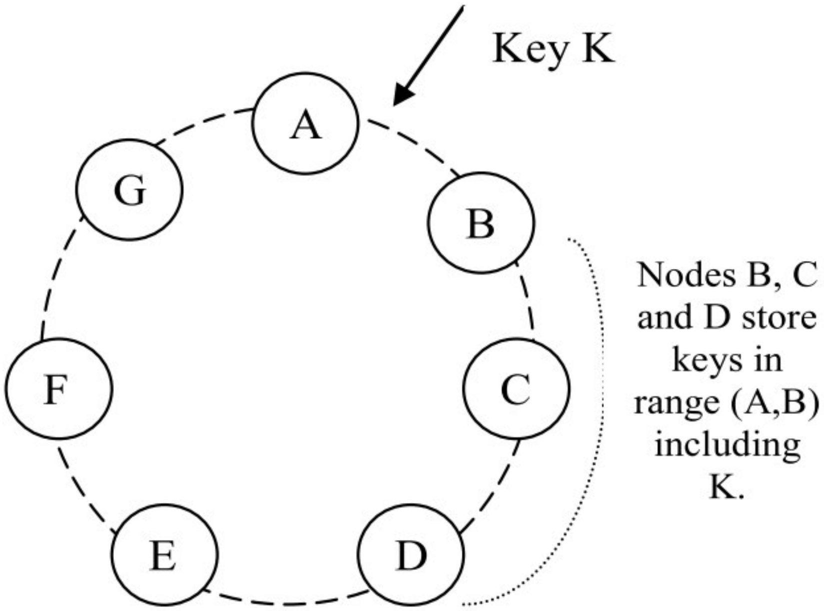
Read and Write Operations

- Coordinator
  - Node responsible for read/writes
  - First node in the preference list
- Write Operation
  - New vector clock from coordinator
  - Write locally and forward to N-1 nodes, if W-1 nodes respond then write was successful
- Read Operation
  - Forward request to N-1 nodes, if R-1 nodes respond then forward to user
  - User resolves conflicts and writes back result

Membership Changes

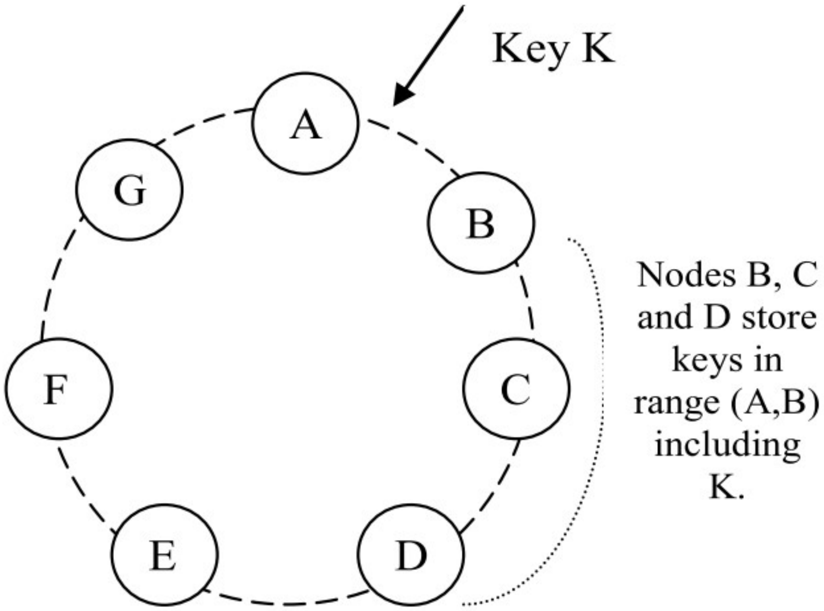
- Gossip-based Protocol to propagate membership changes
  - Each node contacts a peer chosen at random every second and the two nodes efficiently reconcile their persisted membership change histories
- Each node is aware of the key ranges handled by its peers

Handling Failures: Hinted Handoff



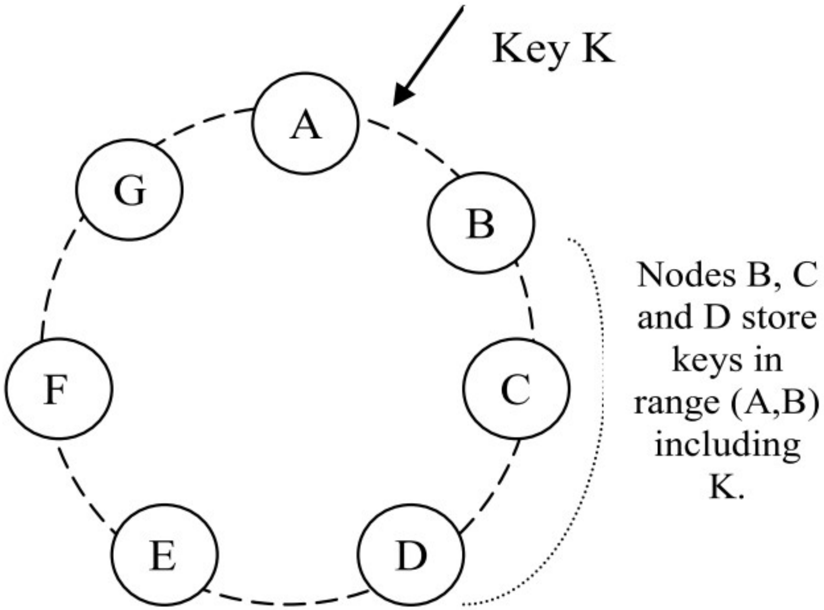
Handling Failures: Hinted Handoff

- Imagine A goes down and N=3



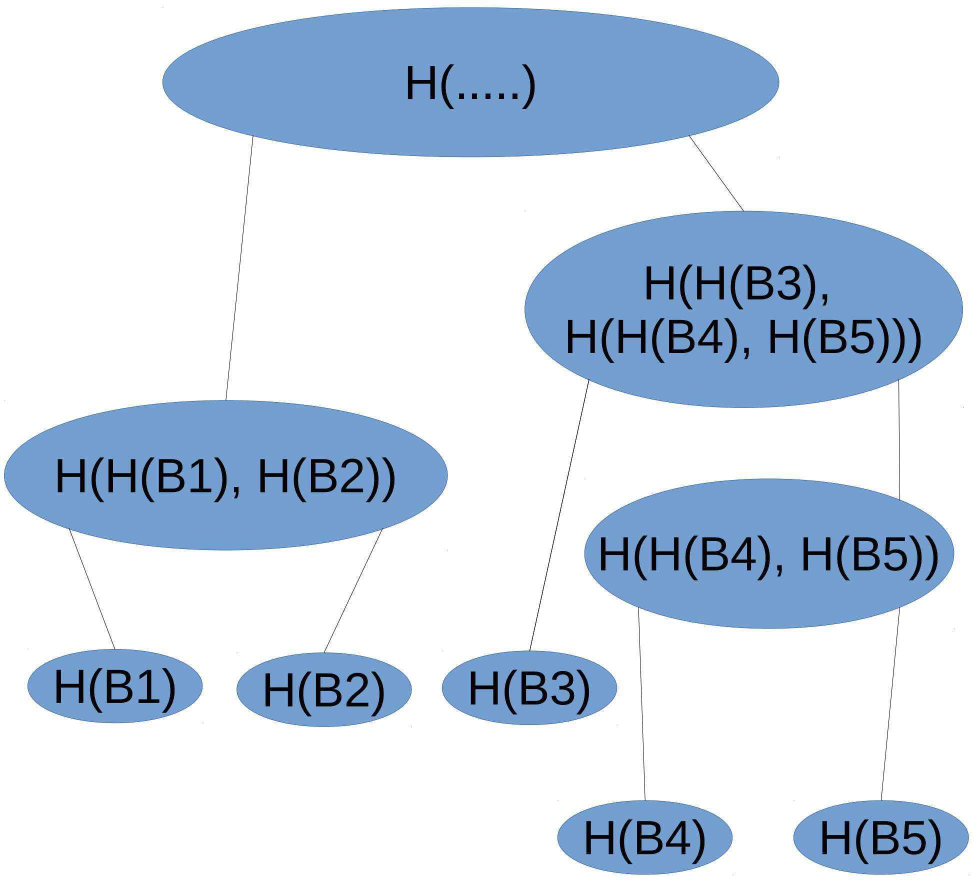
Handling Failures: Hinted Handoff

- Imagine A goes down and N=3
- Keys stored by A will now be stored by D
- D is hinted in the metadata that it is storing keys meant for A
- When A recovers, the keys at D are now copied to A



Handling Failures: Merkle Trees

- Minimize the amount of transferred data
- Merkle Tree:
  - Leaves are hashes of keys
  - Parents are hashes of children
- Each node maintains seperate Merkle tree for each key-range



Summary

Problem	Technique	Advantage
Partitioning	Consistent Hashing	Incremental Scalability
High Availability for writes	Vector clocks with reconciliation during reads	Version size is decoupled from update rates.
Handling temporary failures	Sloppy Quorum and hinted handoff	Provides high availability and durability guarantee when some of the replicas are not available.
Recovering from permanent failures	Anti-entropy using Merkle trees	Synchronizes divergent replicas in the background.
Membership and failure detection	Gossip-based membership protocol and failure detection.	Preserves symmetry and avoids having a centralized registry for storing membership and node liveness information.

G. DeCandia et al. **"Dynamo: Amazon's Highly Available Key-value Store,"** In SOSP 2007.

Modeling Overlay Networks (contd)

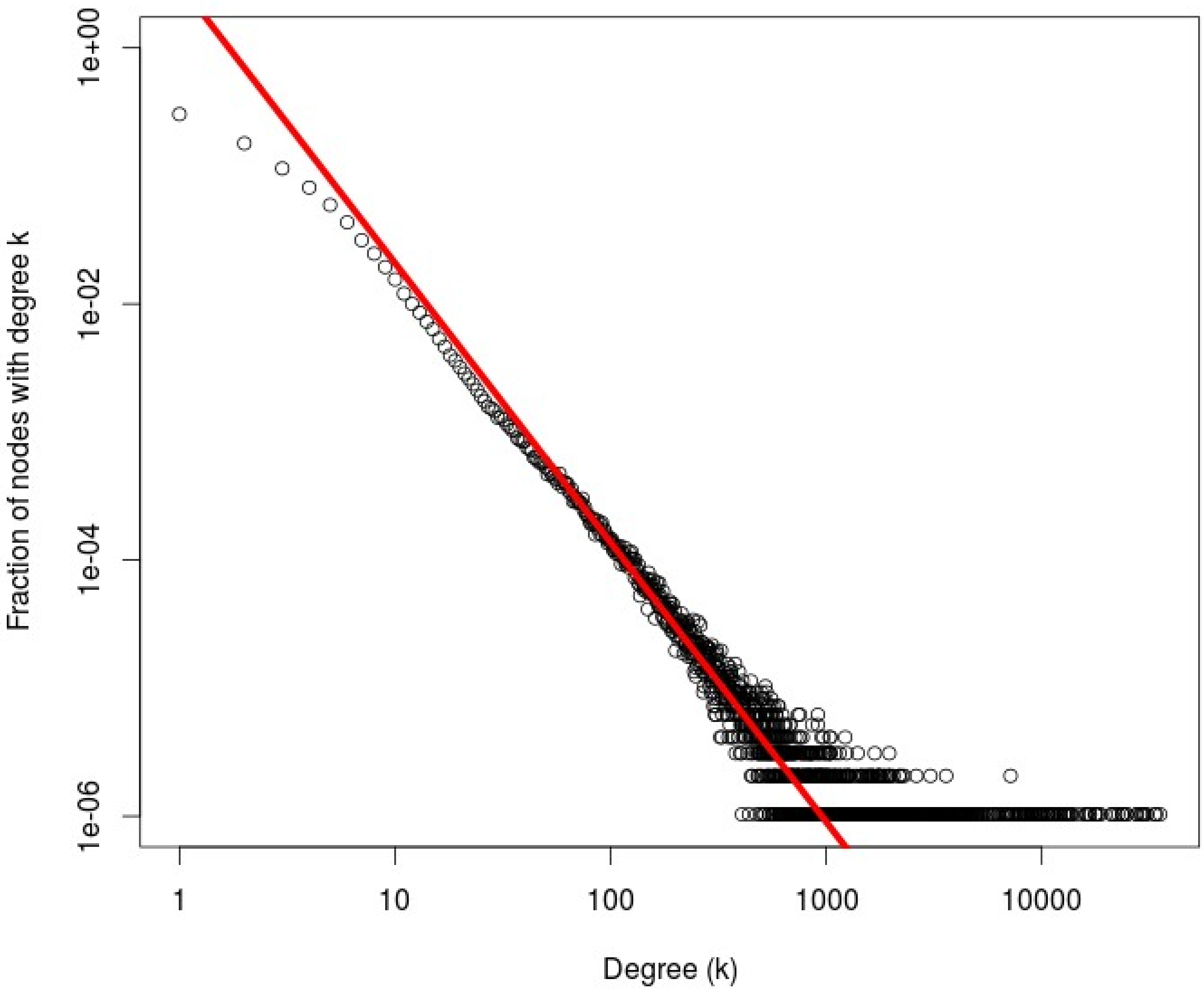
Recap

- Milgram's Experiment
- Duncan Watts Random Rewiring Model
- Scale Free Networks (Power-Law Networks)
  - Preferential attachment
  - Evolving Copying Model (Copying Generative Model)
- Navigation in Small World

Error and Attack Tolerance of Complex Networks

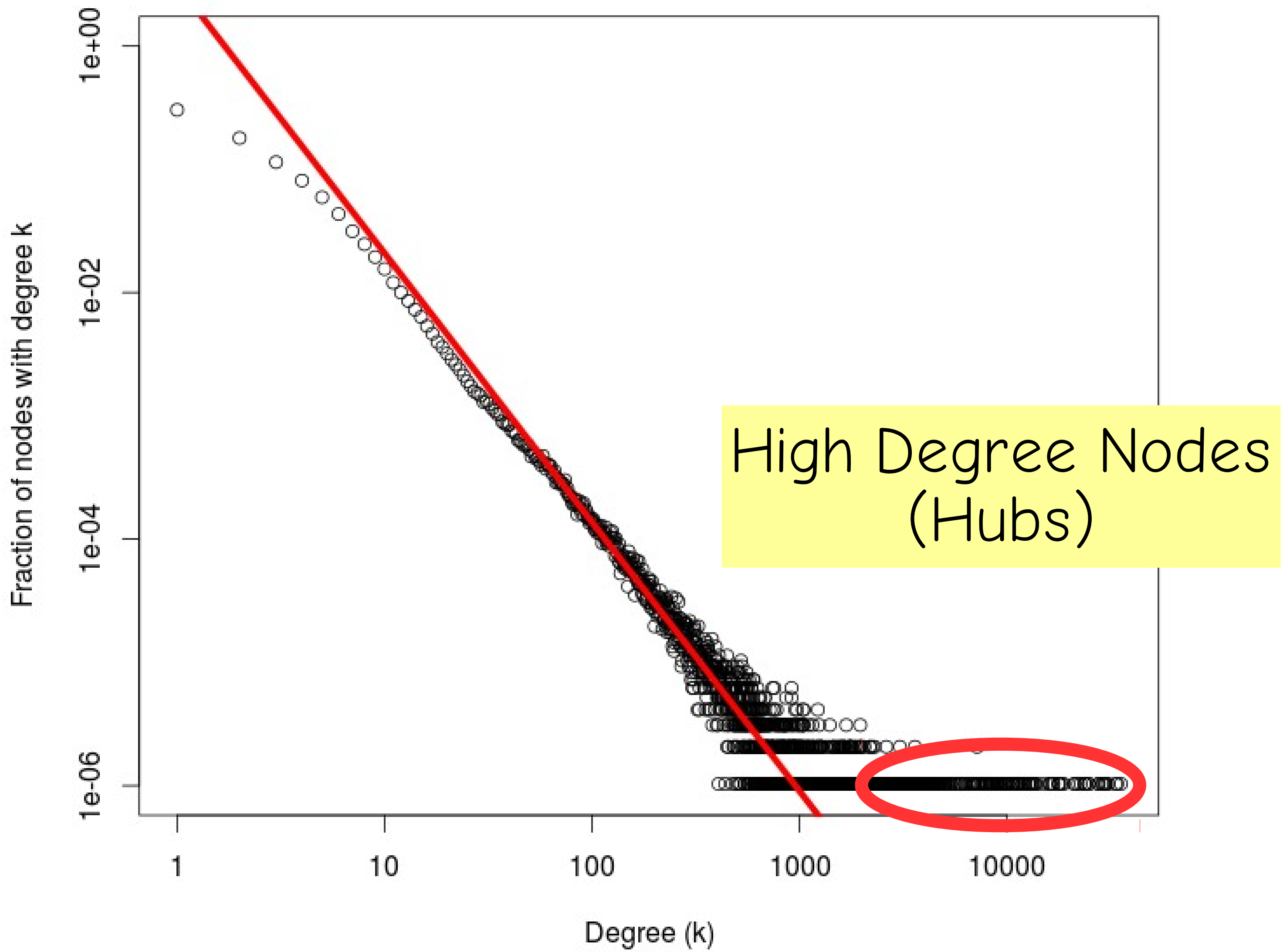
Albert, Réka, et al. **"Error and attack tolerance of complex networks."** nature 406, no. 6794 (2000): 378-382.

Scale-Free Model for AS-Graph



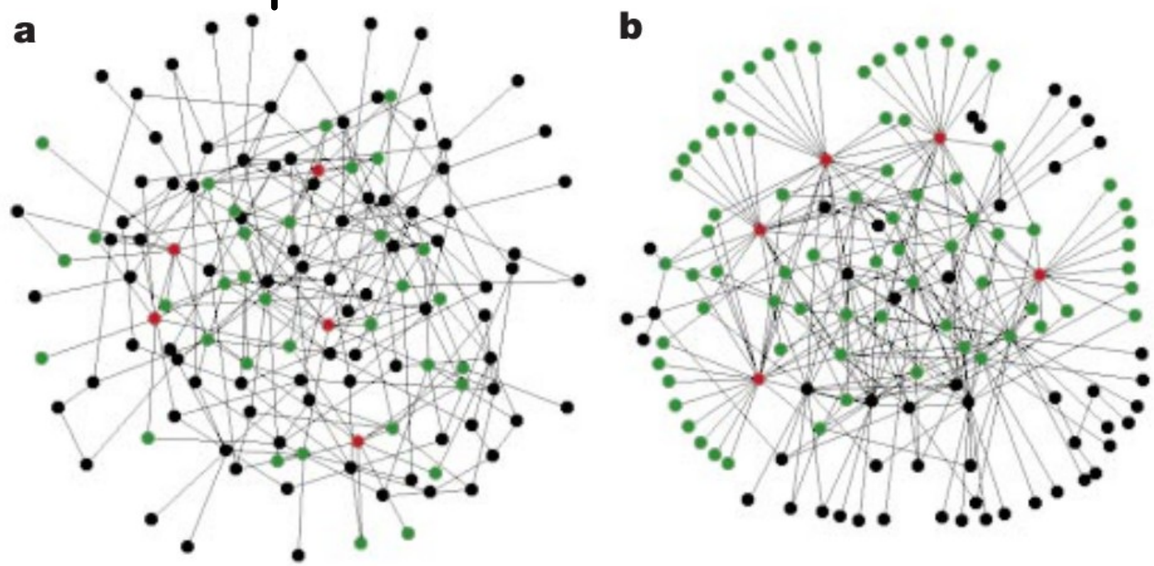
AS Topology of skitter dataset parsed by SNAP team
http://snap.stanford.edu/data/as-skitter.html

Scale-Free Model for AS-Graph



AS Topology of skitter dataset parsed by SNAP team
http://snap.stanford.edu/data/as-skitter.html

Importance of Hubs



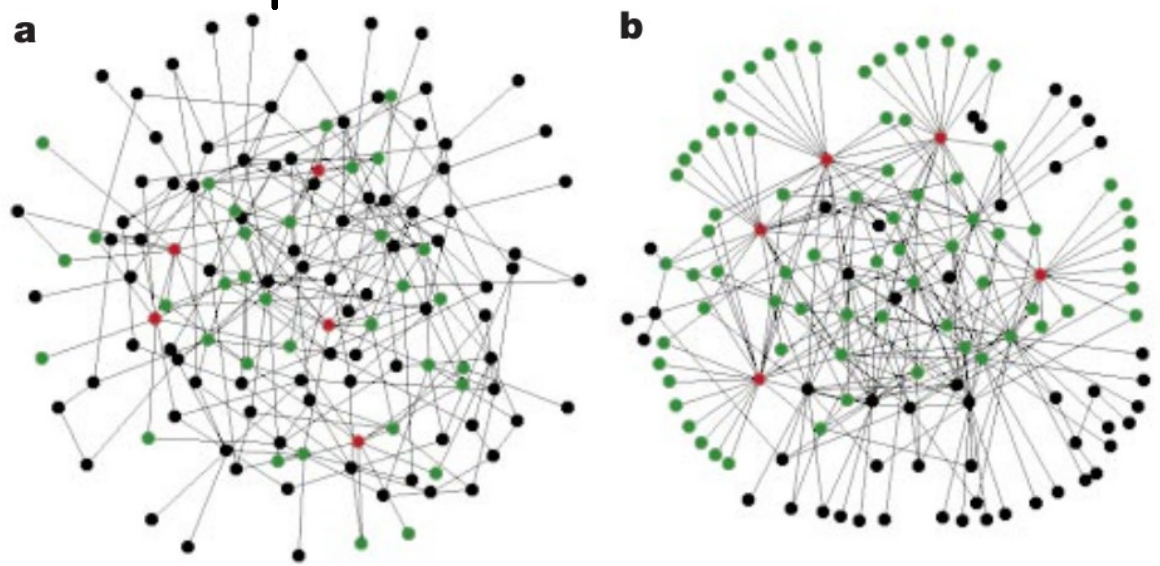
(Random Graph) Exponential

Scale-free

Figure 1 Visual illustration of the difference between an exponential and a scale-free network. **a**, The exponential network is homogeneous: most nodes have approximately the same number of links. **b**, The scale-free network is inhomogeneous: the majority of the nodes have one or two links but a few nodes have a large number of links, guaranteeing that the system is fully connected. Red, the five nodes with the highest number of links; green, their first neighbours. Although in the exponential network only 27% of the nodes are reached by the five most connected nodes, in the scale-free network more than 60% are reached, demonstrating the importance of the connected nodes in the scale-free network Both networks contain 130 nodes and 215 links ($\langle k \rangle = 3.3$). The network visualization was done using the Pajek program for large network analysis: ⟨http://vlado.fmf.uni-lj.si/pub/networks/pajek/pajekman.htm⟩.

Albert, Réka, et al. **"Error and attack tolerance of complex networks."** nature 406, no. 6794 (2000): 378-382.

Importance of Hubs

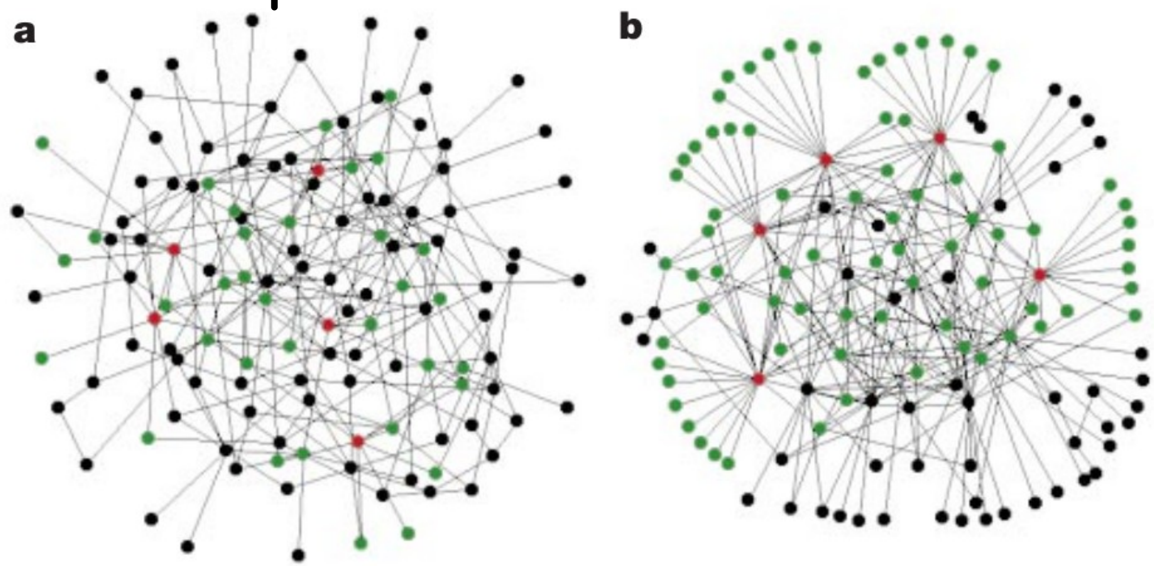


(Random Graph) Exponential Scale-free

Figure 1 Visual illustration of the difference between an exponential and a scale-free network. **a**, The exponential network is homogeneous: most nodes have approximately the same number of links. **b**, The scale-free network is inhomogeneous: the majority of the nodes have one or two links but a few nodes have a large number of links, guaranteeing that the system is fully connected. Red, the five nodes with the highest number of links; green, their first neighbours. Although in the exponential network only 27% of the nodes are reached by the five most connected nodes, in the scale-free network more than 60% are reached, demonstrating the importance of the connected nodes in the scale-free network Both networks contain 130 nodes and 215 links ($\langle k \rangle = 3.3$). The network visualization was done using the Pajek program for large network analysis: ⟨http://vlado.fmf.uni-lj.si/pub/networks/pajek/pajekman.htm⟩.

Albert, Réka, et al. **"Error and attack tolerance of complex networks."** nature 406, no. 6794 (2000): 378-382.

Importance of Hubs



(Random Graph) Exponential Scale-free

Figure 1 Visual illustration of the difference between an exponential and a scale-free network. **a**, The exponential network is homogeneous: most nodes have approximately the same number of links. **b**, The scale-free network is inhomogeneous: the majority of the nodes have one or two links but a few nodes have a large number of links, guaranteeing that the system is fully connected. Red, the five nodes with the highest number of links; green, their first neighbours. Although in the exponential network only 27% of the nodes are reached by the five most connected nodes, in the scale-free network more than 60% are reached, demonstrating the importance of the connected nodes in the scale-free network Both networks contain 130 nodes and 215 links ($\langle k \rangle = 3.3$). The network visualization was done using the Pajek program for large network analysis: ⟨http://vlado.fmf.uni-lj.si/pub/networks/pajek/pajekman.htm⟩.

Albert, Réka, et al. **"Error and attack tolerance of complex networks."** nature 406, no. 6794 (2000): 378-382.

Error vs Attack

- Error (Node Failure)

Error vs Attack

- Error (Node Failure)
  - random node fails (malfunction)

Error vs Attack

- Error (Node Failure)
  - random node fails (malfunction)
- Attack

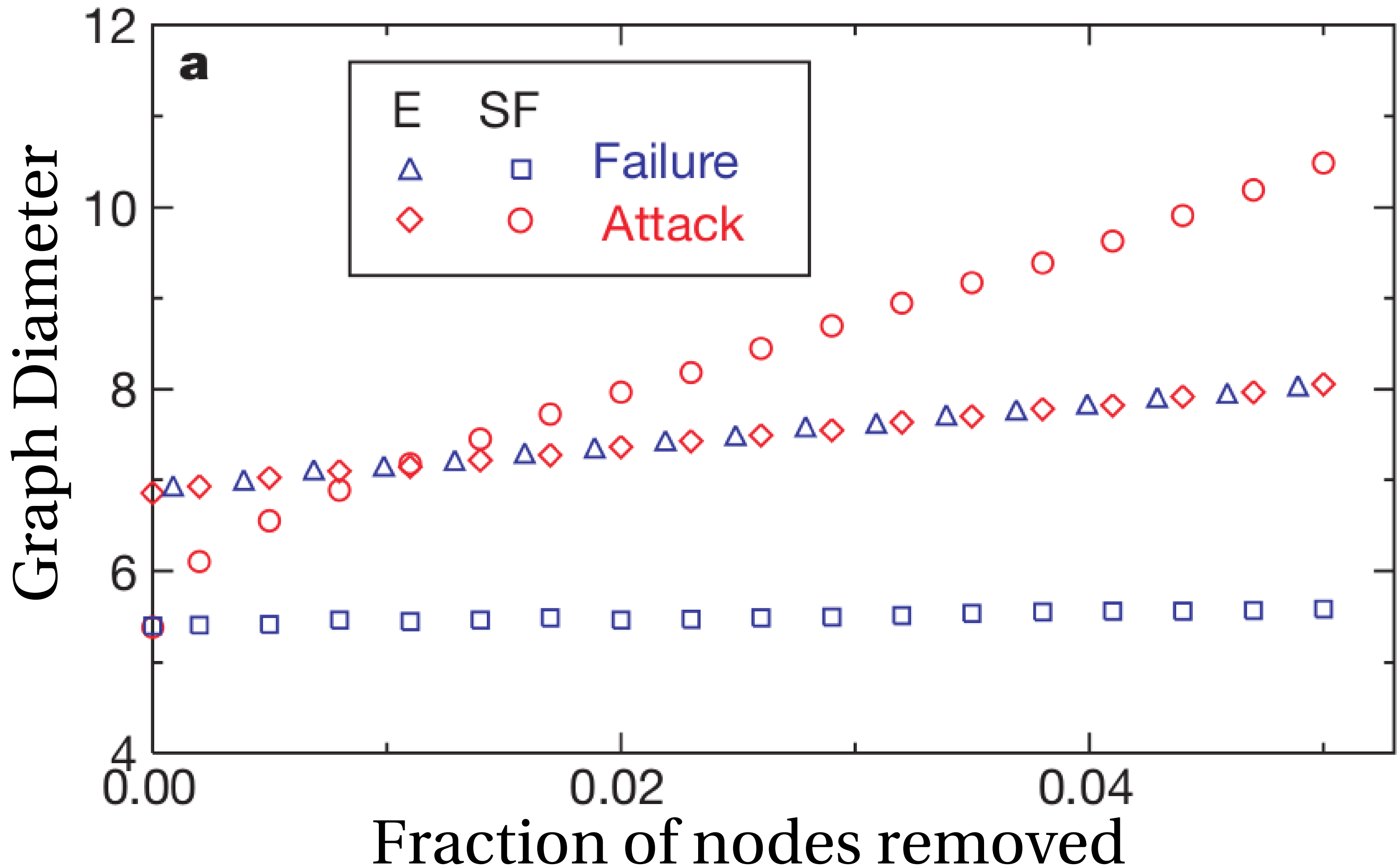
Error vs Attack

- Error (Node Failure)
  - random node fails (malfunction)
- Attack
  - Selected node with a given property is made to fail
  - Which nodes would you target if you knew the network is a scale-free network?

Error vs Attack

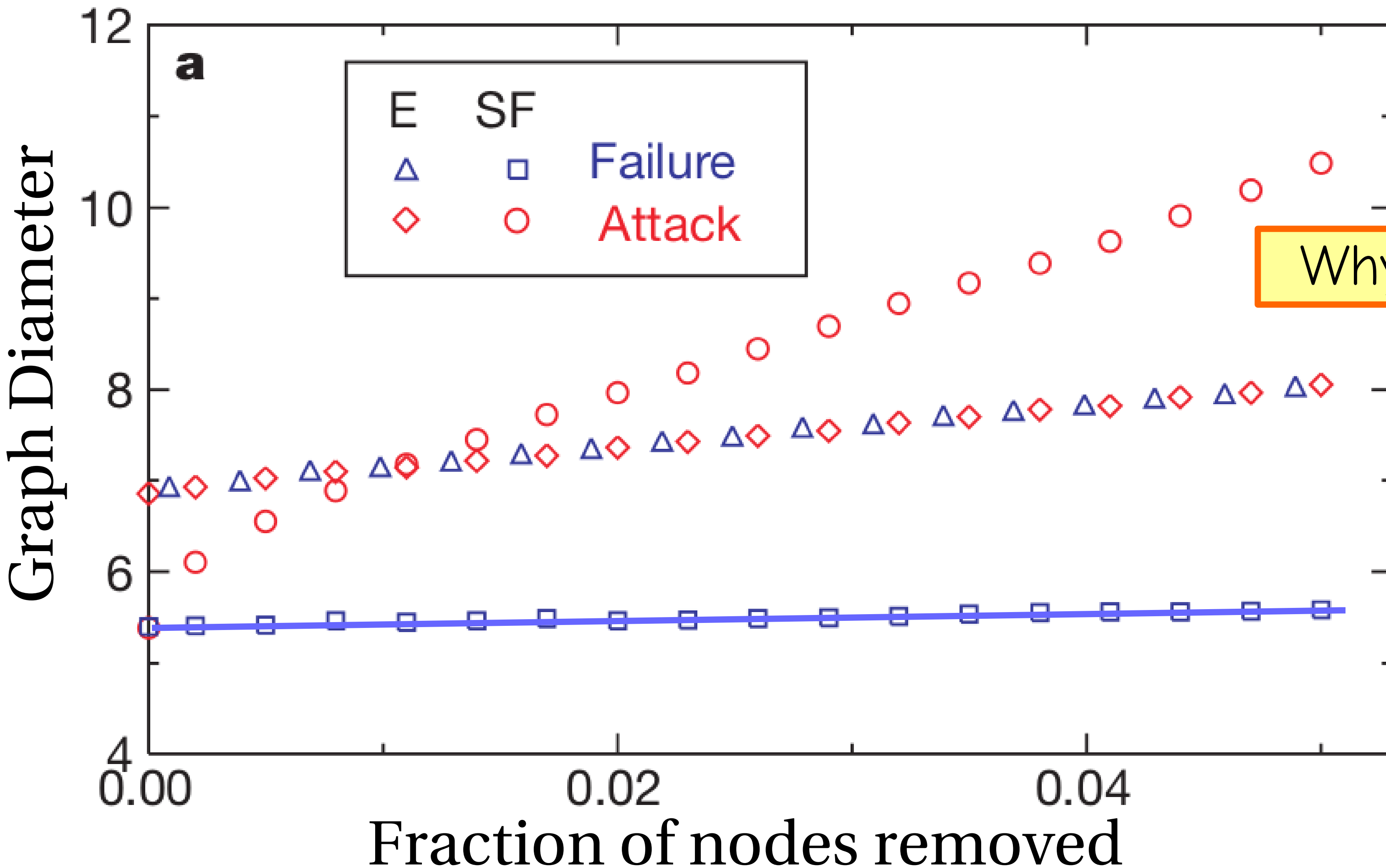
- Error (Node Failure)
  – random node fails (malfunction)
- Attack
  – Selected node with a given property is made to fail
  – Which nodes would you target if you knew the network is a scale-free network?
    - ***Nodes with the highest degree***

Impact of Errors and Attacks (Graph Diameter)



Albert, Réka, et al. **"Error and attack tolerance of complex networks."** nature 406, no. 6794 (2000): 378-382.

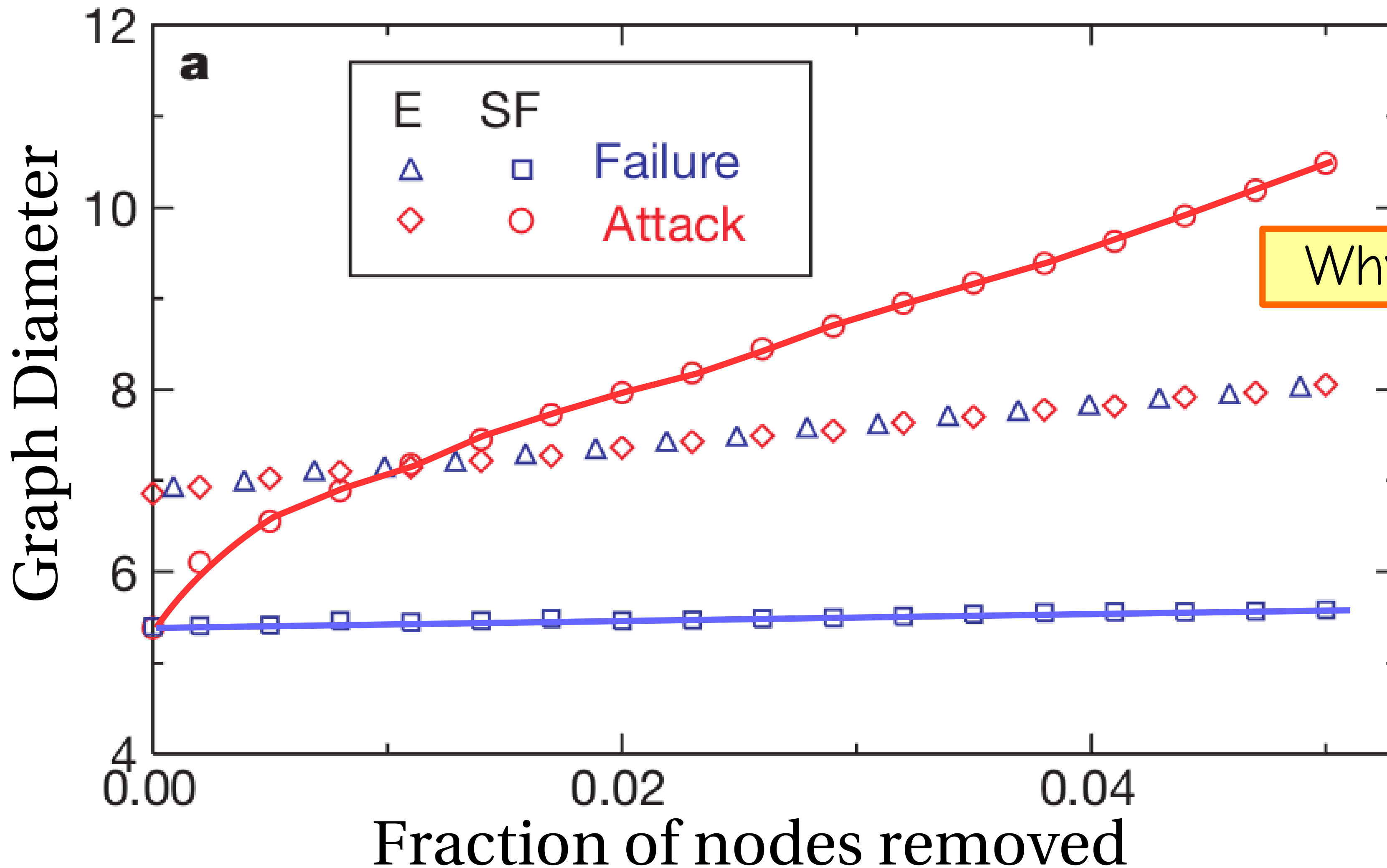
Impact of Errors and Attacks (Graph Diameter)



Why?

Albert, Réka, et al. **"Error and attack tolerance of complex networks."** nature 406, no. 6794 (2000): 378-382.

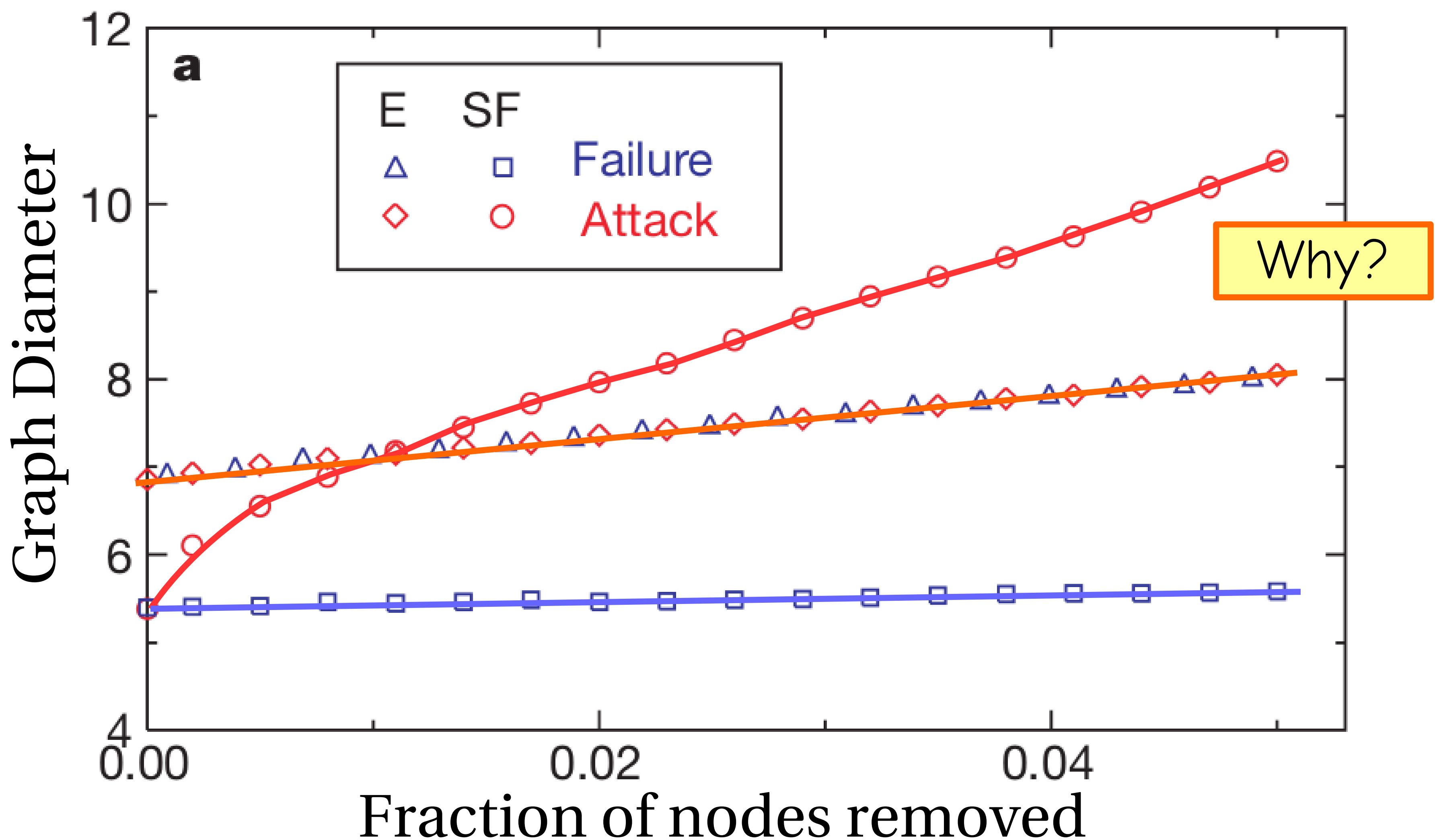
Impact of Errors and Attacks (Graph Diameter)



Why?

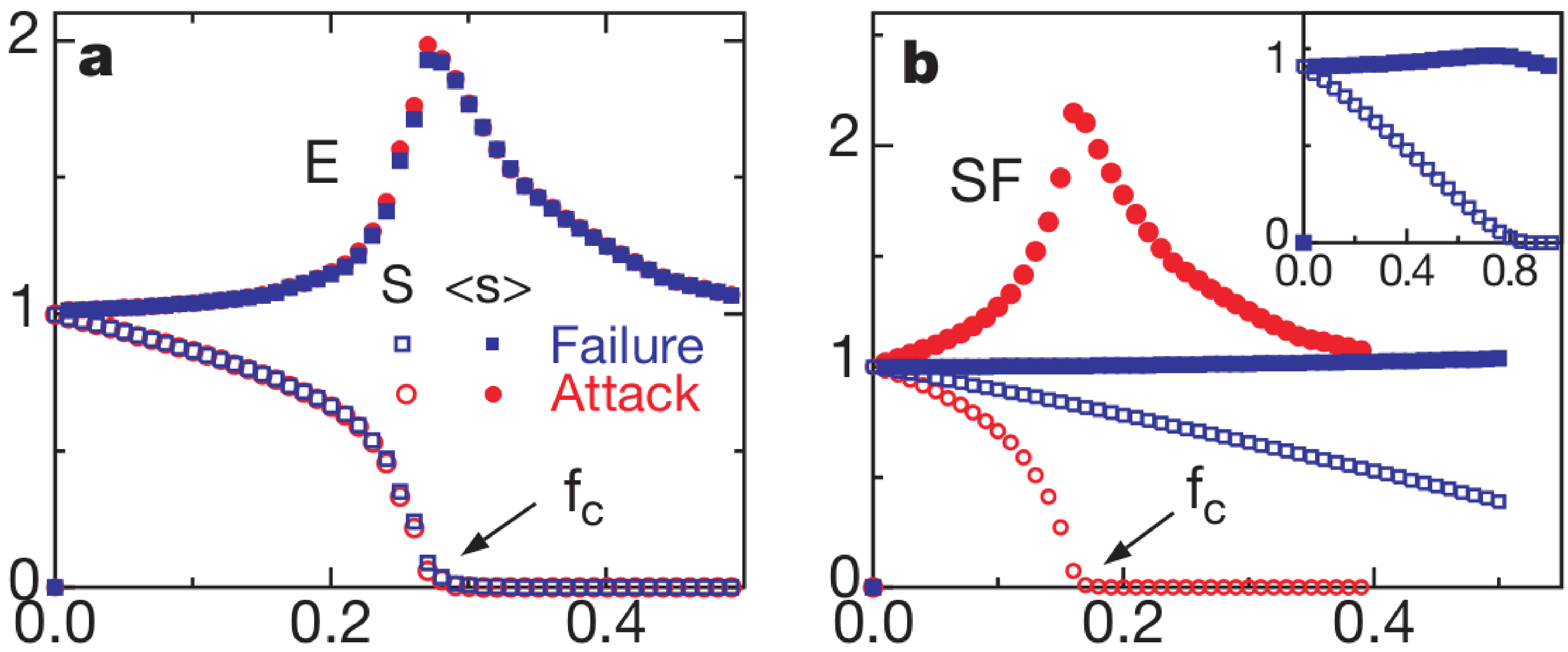
Albert, Réka, et al. **"Error and attack tolerance of complex networks."** nature 406, no. 6794 (2000): 378-382.

Impact of Errors and Attacks (Graph Diameter)



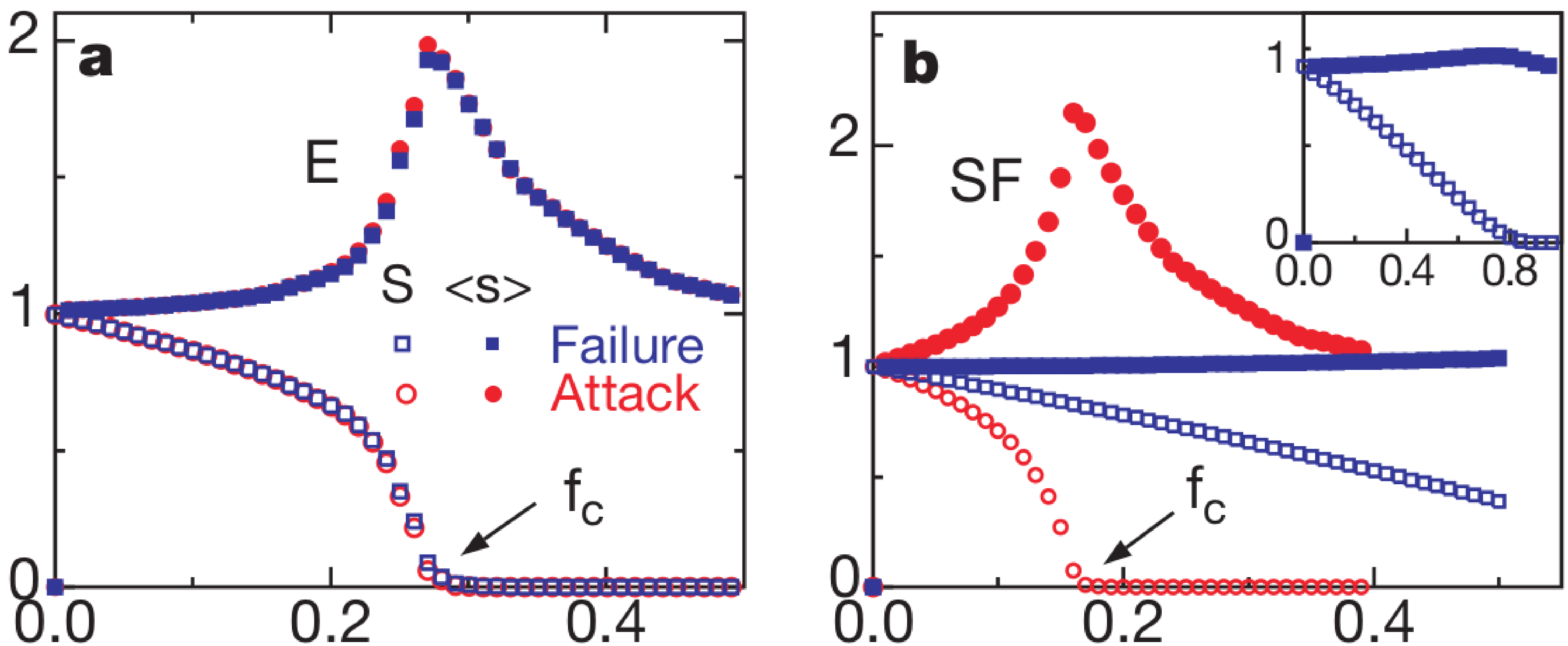
Albert, Réka, et al. **"Error and attack tolerance of complex networks."** nature 406, no. 6794 (2000): 378-382.

Impact of Errors and Attacks (Size of Largest Cluster)



Albert, Réka, et al. **"Error and attack tolerance of complex networks."** nature 406, no. 6794 (2000): 378-382.

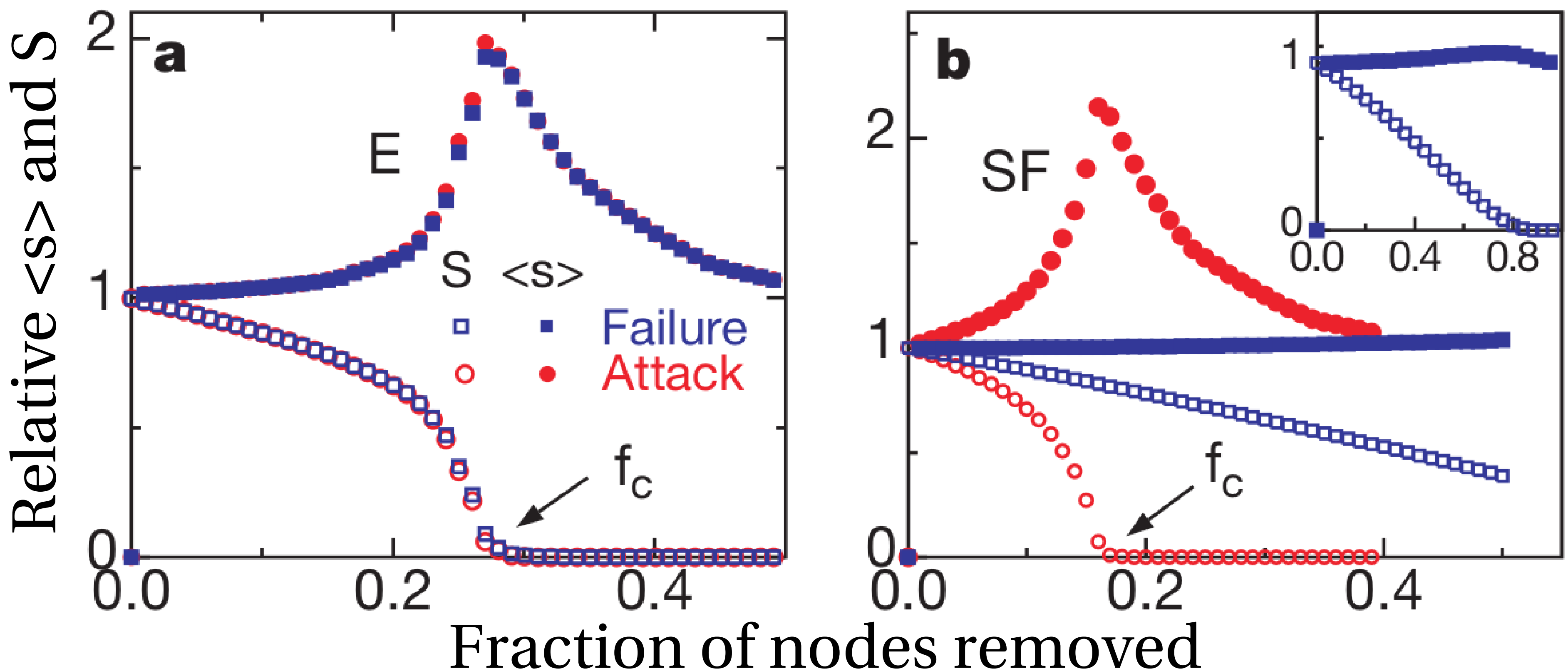
Impact of Errors and Attacks (Size of Largest Cluster)



Fraction of nodes removed

Albert, Réka, et al. **"Error and attack tolerance of complex networks."** nature 406, no. 6794 (2000): 378-382.

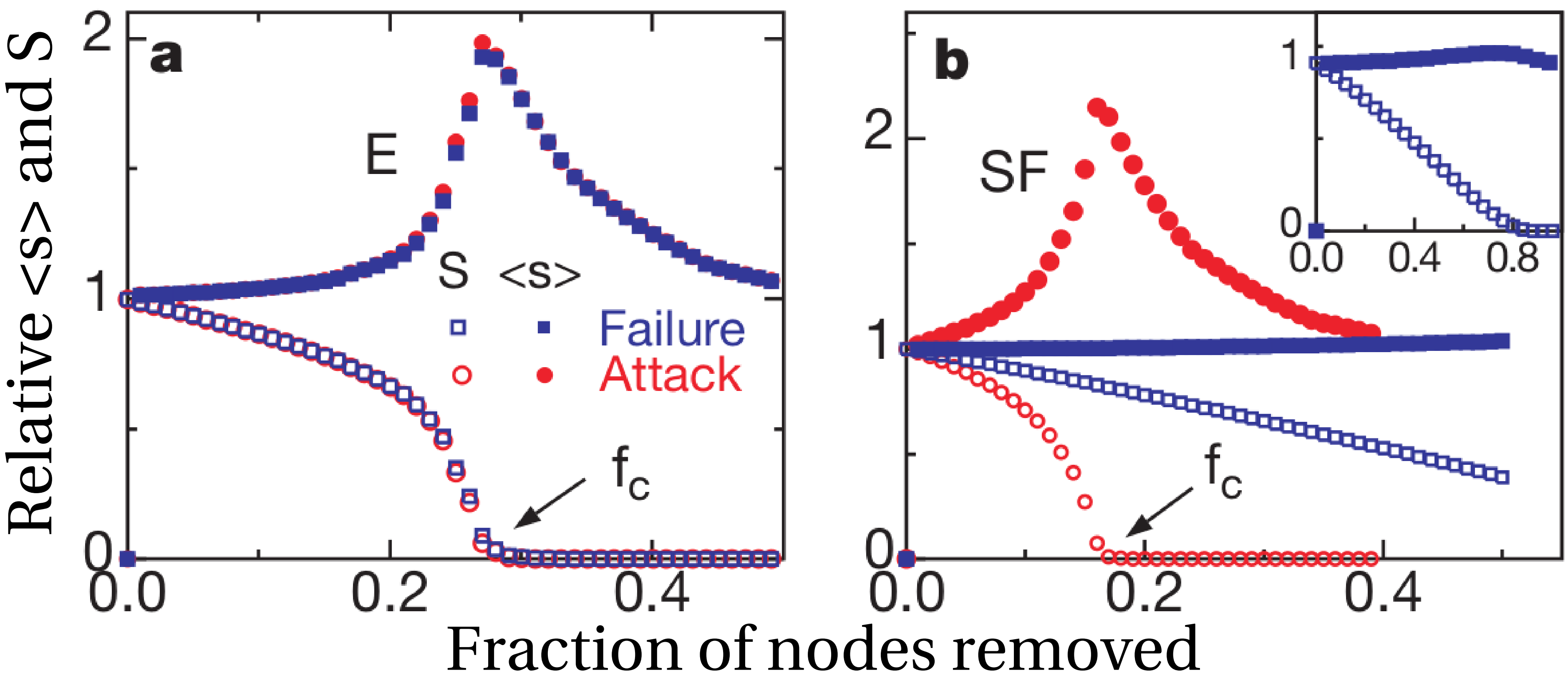
Impact of Errors and Attacks (Size of Largest Cluster)



Albert, Réka, et al. **"Error and attack tolerance of complex networks."** nature 406, no. 6794 (2000): 378-382.

Impact of Errors and Attacks (Size of Largest Cluster)

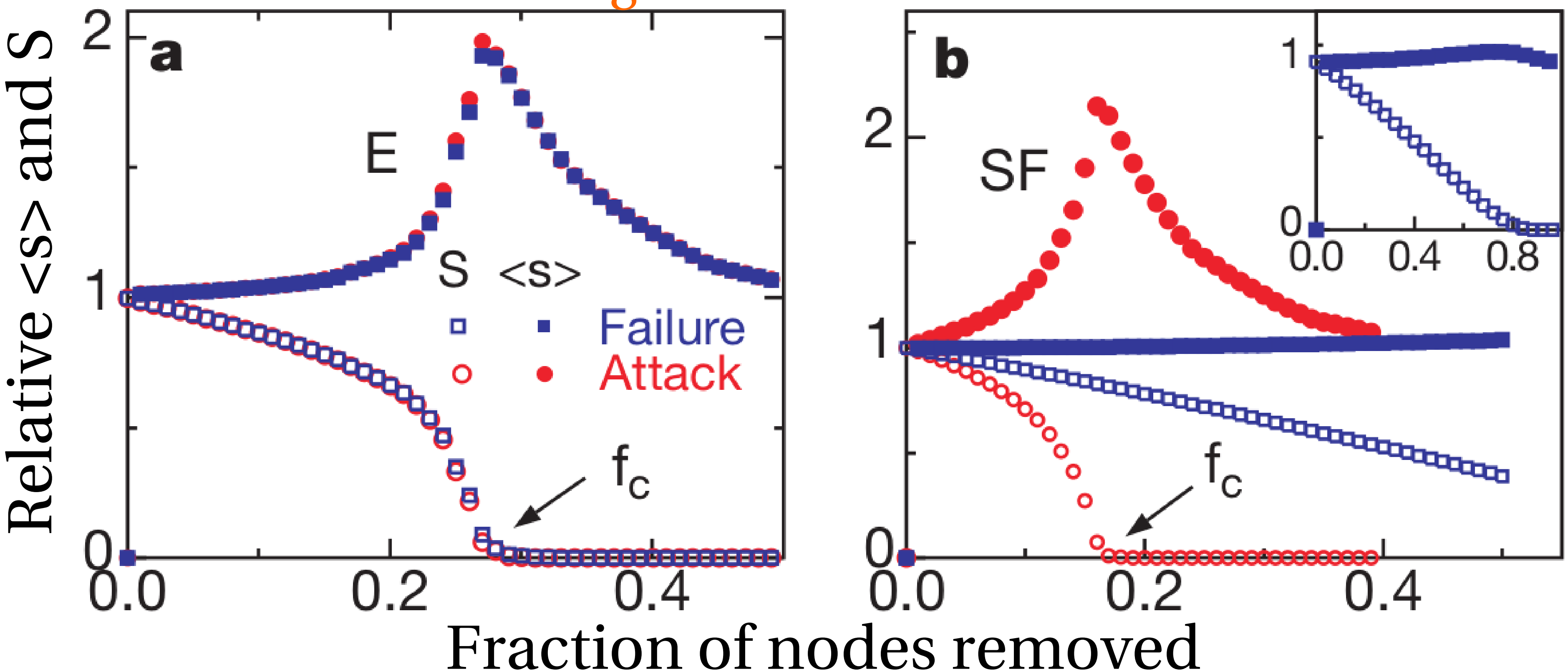
S: Fraction of nodes in largest cluster



Albert, Réka, et al. **"Error and attack tolerance of complex networks."** nature 406, no. 6794 (2000): 378-382.

Impact of Errors and Attacks (Size of Largest Cluster)

S: Fraction of nodes in largest cluster
<s>: average size of isolated clusters

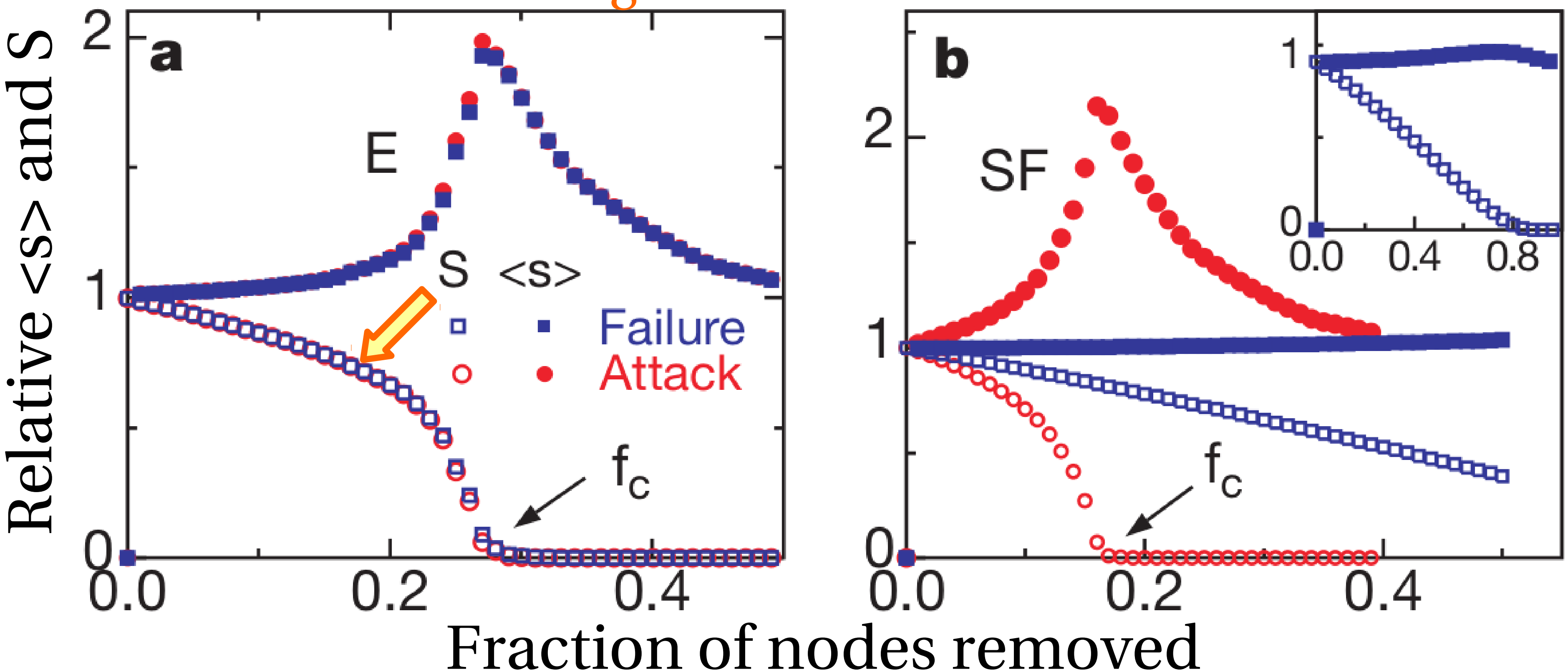


Albert, Réka, et al. **"Error and attack tolerance of complex networks."** nature 406, no. 6794 (2000): 378-382.

Impact of Errors and Attacks (Size of Largest Cluster)

S: Fraction of nodes in largest cluster

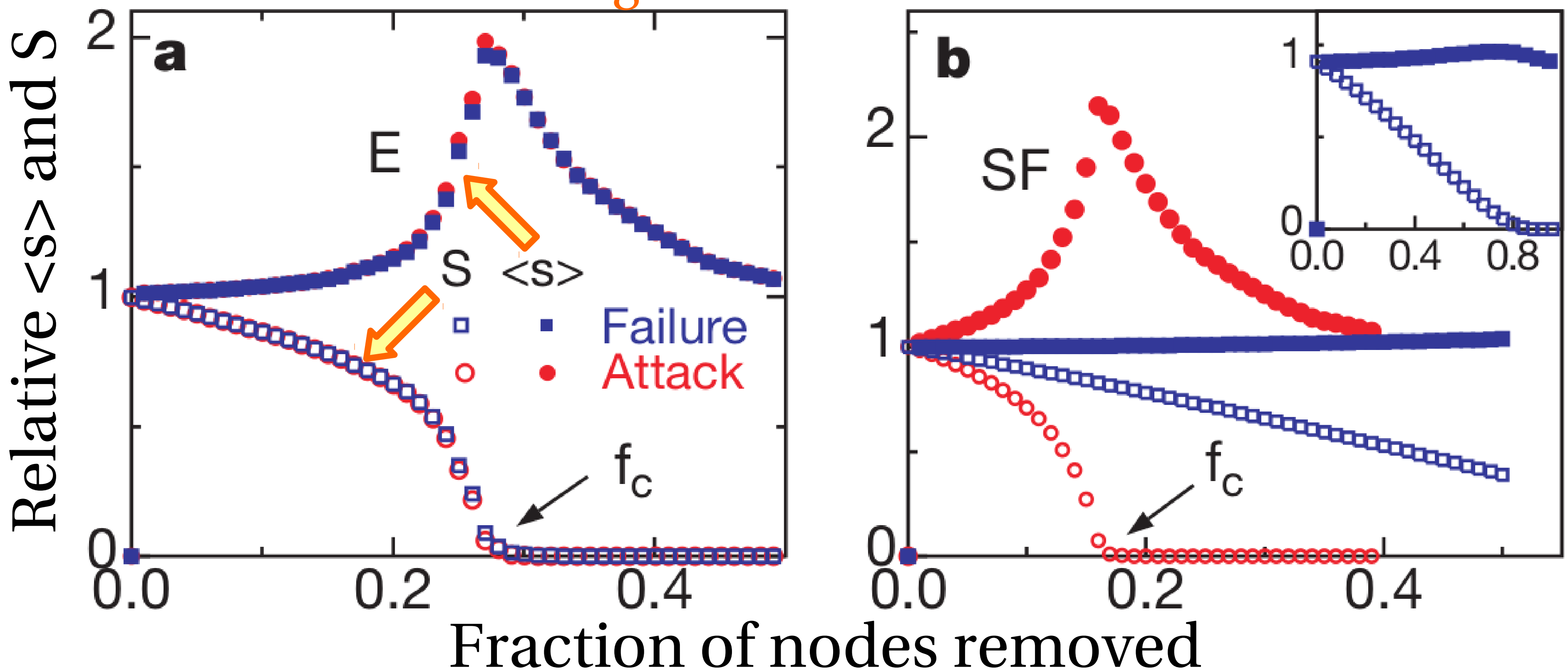
<s>: average size of isolated clusters



Albert, Réka, et al. **"Error and attack tolerance of complex networks."** nature 406, no. 6794 (2000): 378-382.

Impact of Errors and Attacks (Size of Largest Cluster)

S: Fraction of nodes in largest cluster
<s>: average size of isolated clusters

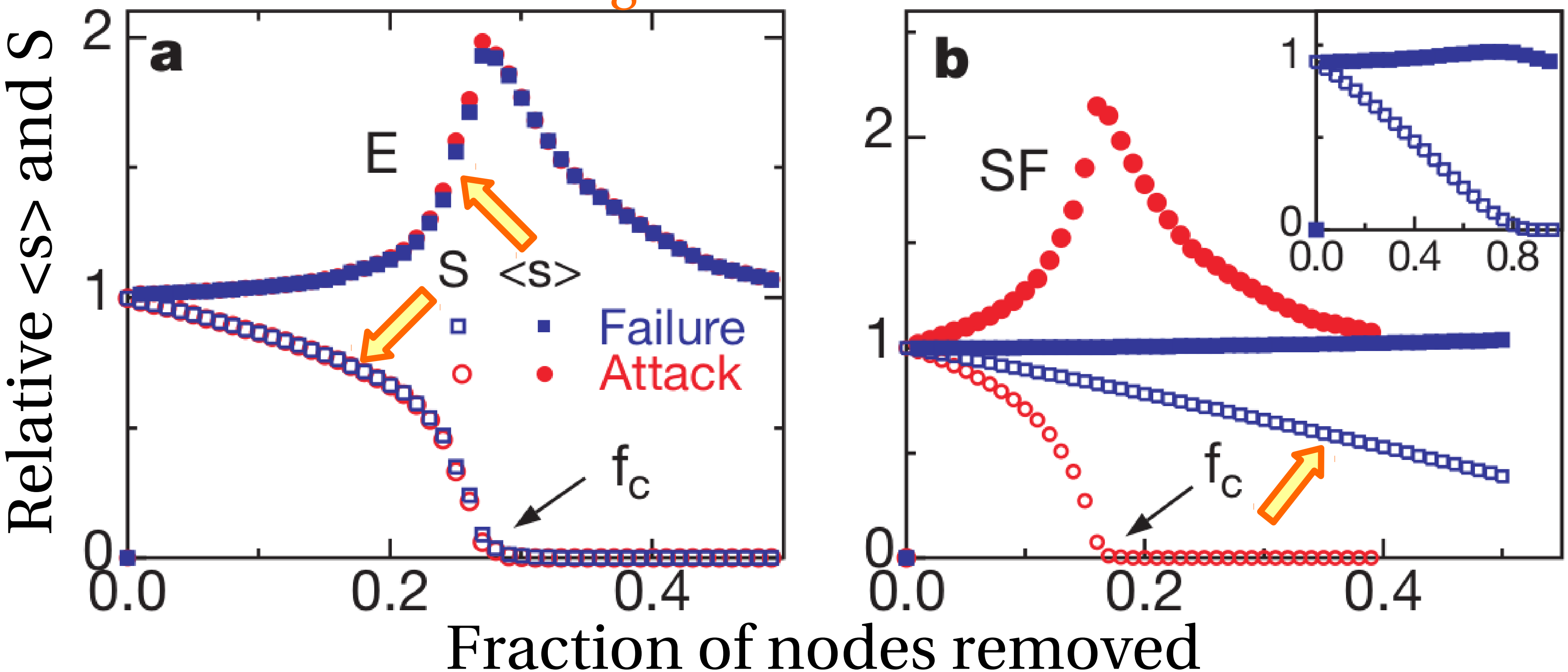


Albert, Réka, et al. **"Error and attack tolerance of complex networks."** nature 406, no. 6794 (2000): 378-382.

Impact of Errors and Attacks (Size of Largest Cluster)

S: Fraction of nodes in largest cluster

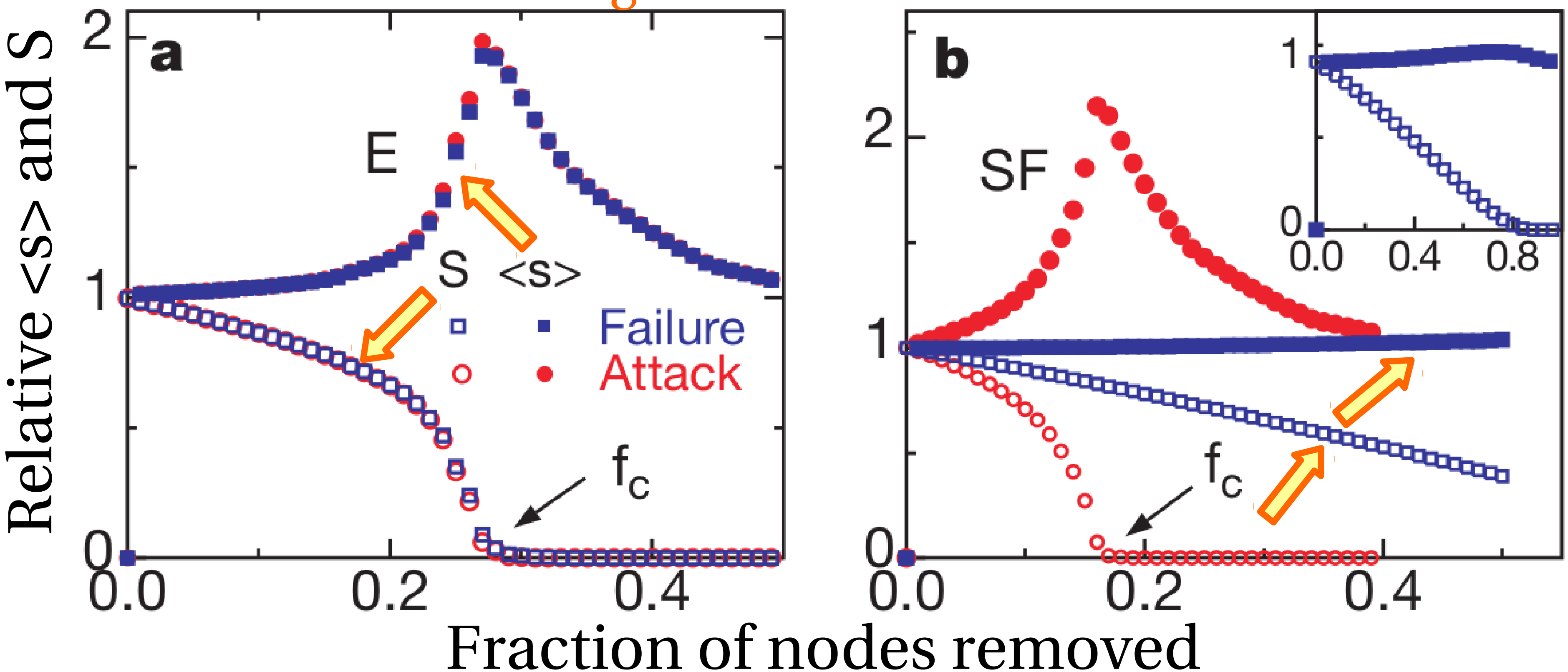
<s>: average size of isolated clusters



Albert, Réka, et al. **"Error and attack tolerance of complex networks."** nature 406, no. 6794 (2000): 378-382.

Impact of Errors and Attacks (Size of Largest Cluster)

S: Fraction of nodes in largest cluster
<s>: average size of isolated clusters

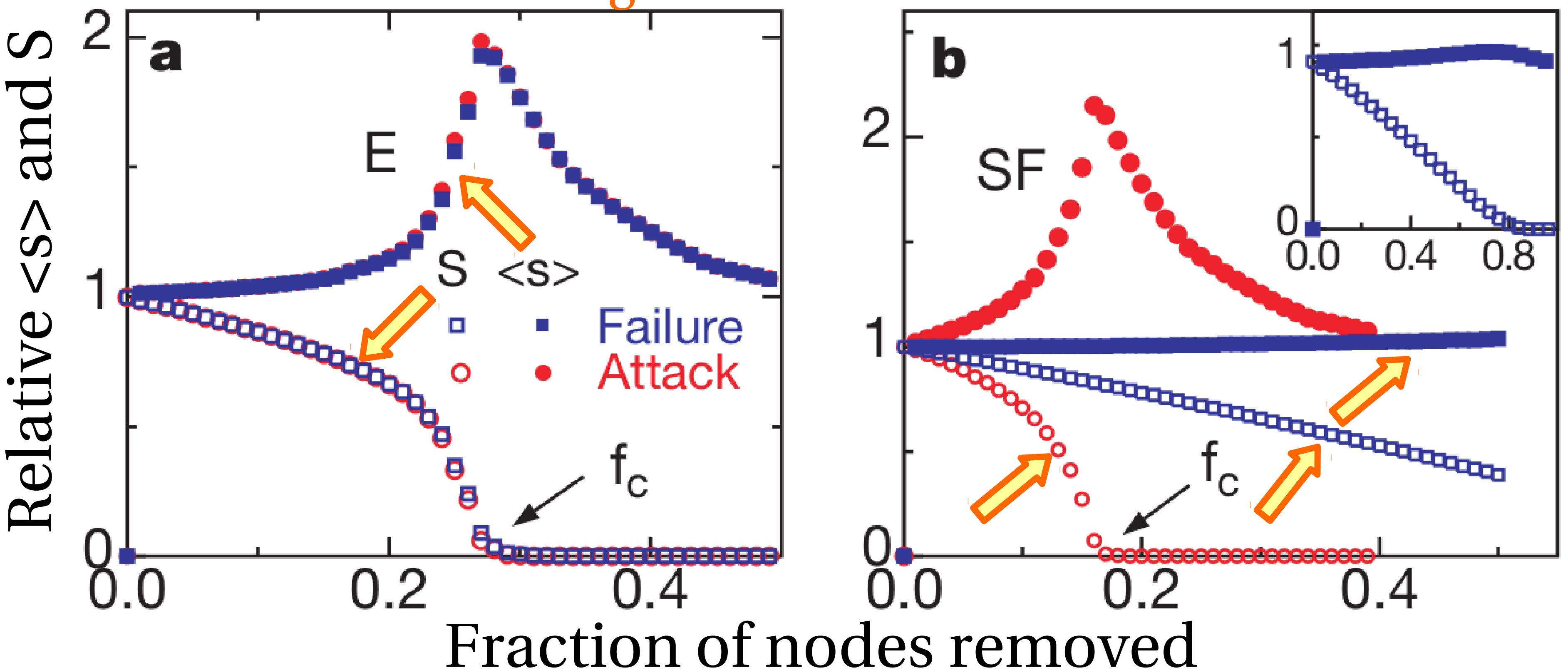


Albert, Réka, et al. **"Error and attack tolerance of complex networks."** nature 406, no. 6794 (2000): 378-382.

Impact of Errors and Attacks (Size of Largest Cluster)

S: Fraction of nodes in largest cluster

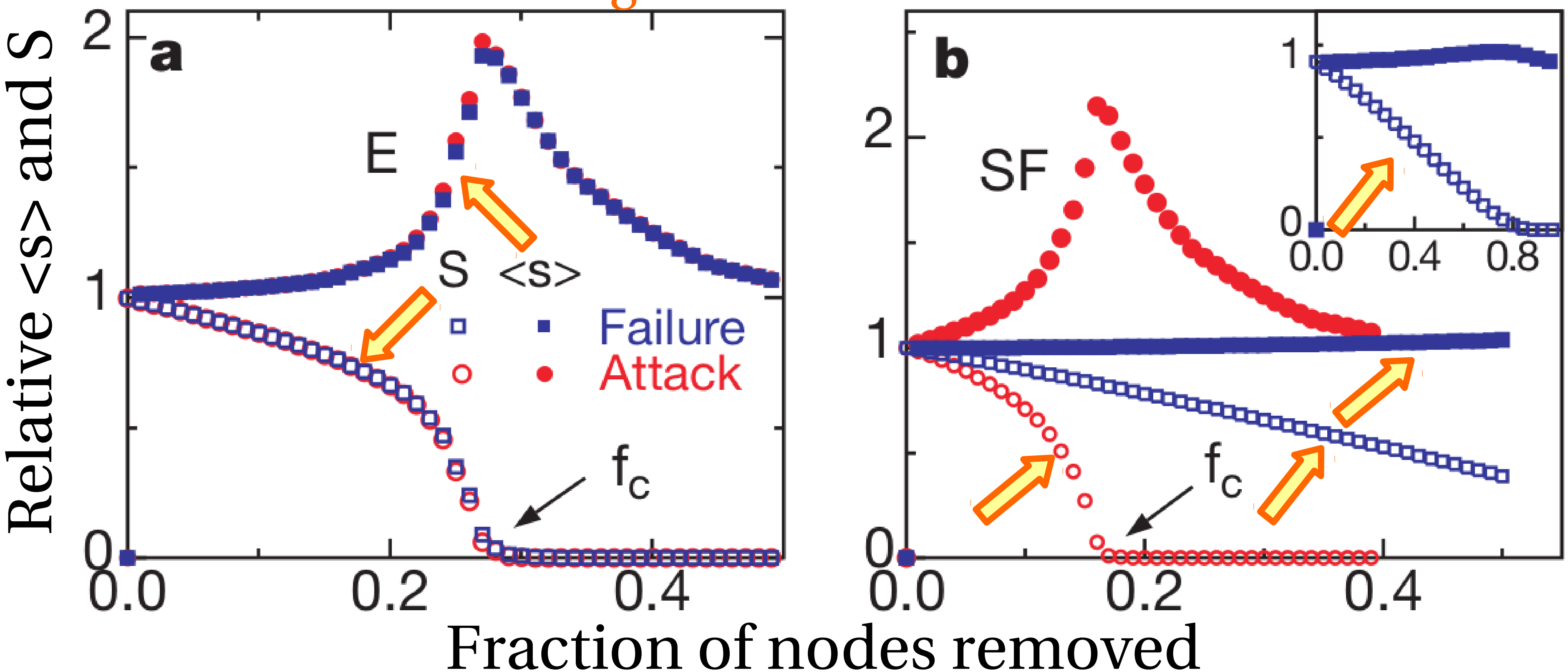
<s>: average size of isolated clusters



Albert, Réka, et al. **"Error and attack tolerance of complex networks."** nature 406, no. 6794 (2000): 378-382.

Impact of Errors and Attacks (Size of Largest Cluster)

S: Fraction of nodes in largest cluster
<s>: average size of isolated clusters

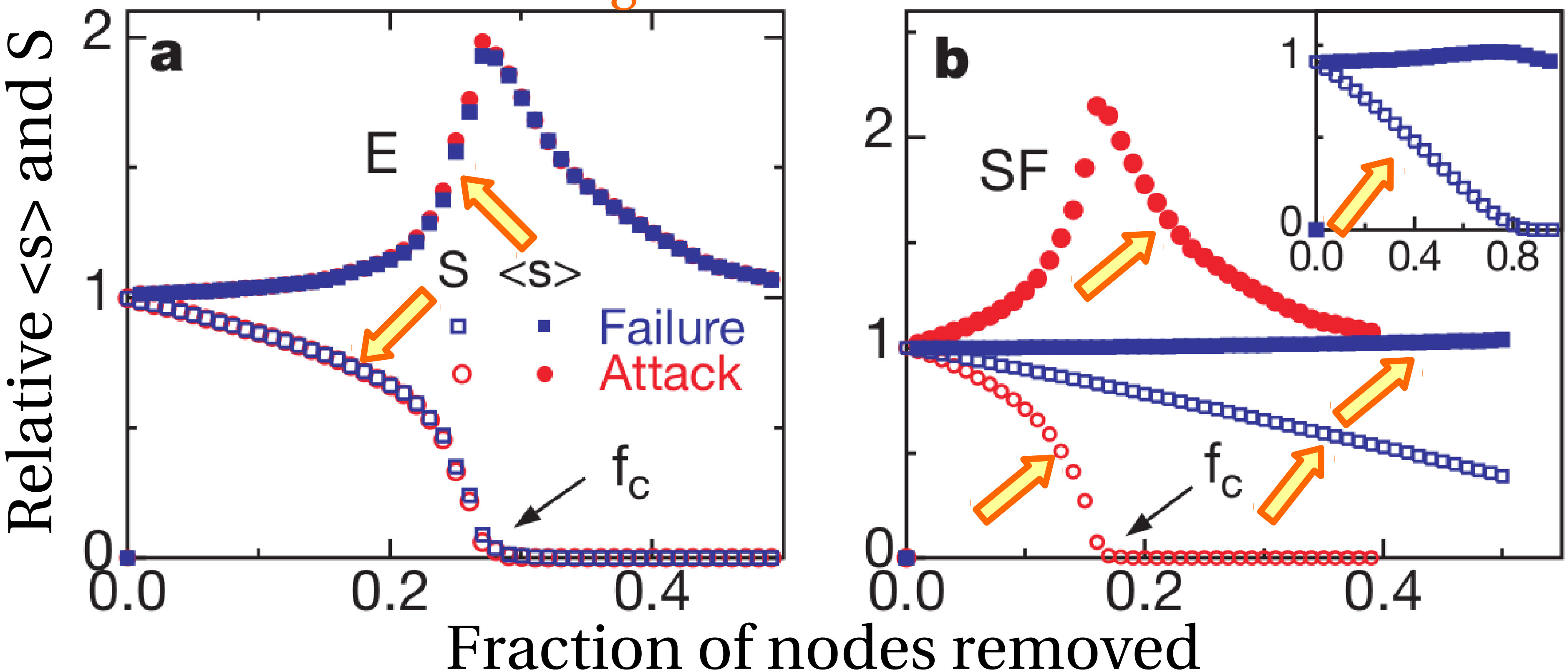


Albert, Réka, et al. **"Error and attack tolerance of complex networks."** nature 406, no. 6794 (2000): 378-382.

Impact of Errors and Attacks (Size of Largest Cluster)

S: Fraction of nodes in largest cluster

<s>: average size of isolated clusters

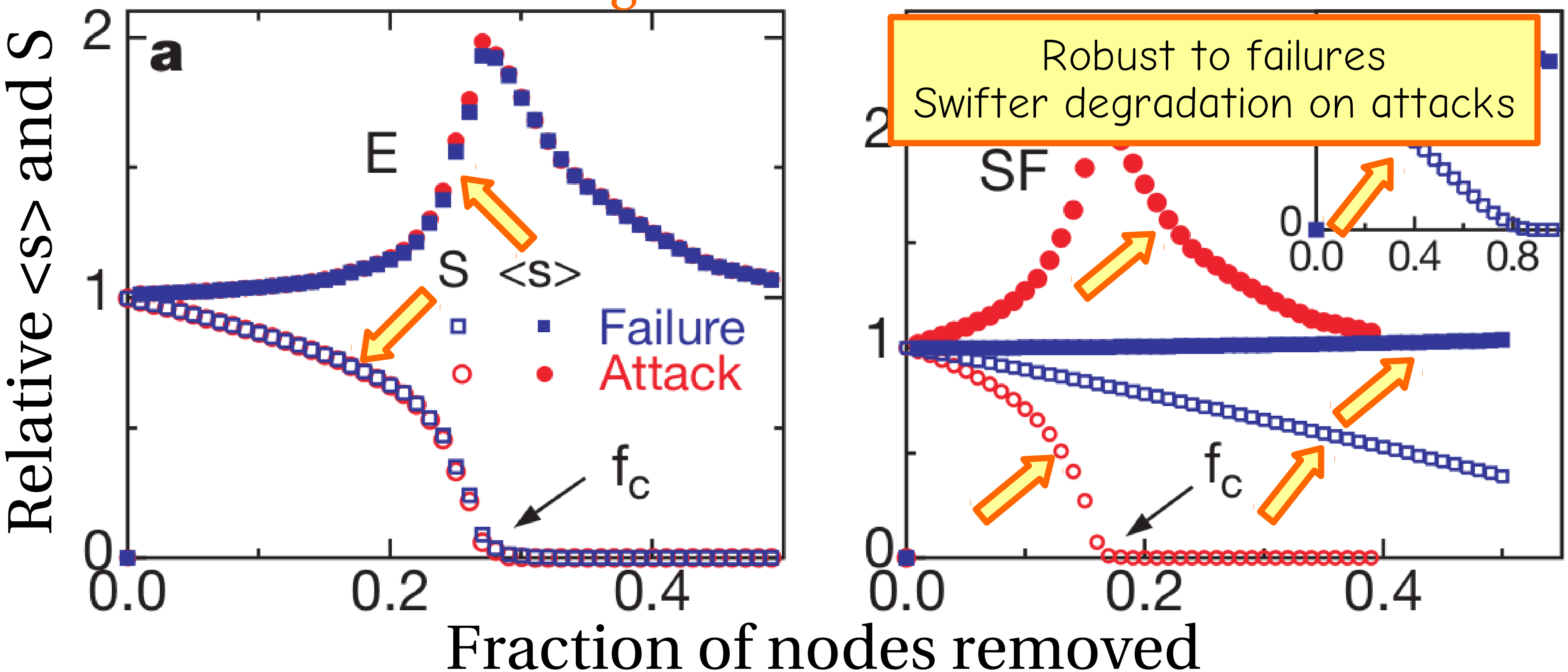


Albert, Réka, et al. **"Error and attack tolerance of complex networks."** nature 406, no. 6794 (2000): 378-382.

Impact of Errors and Attacks (Size of Largest Cluster)

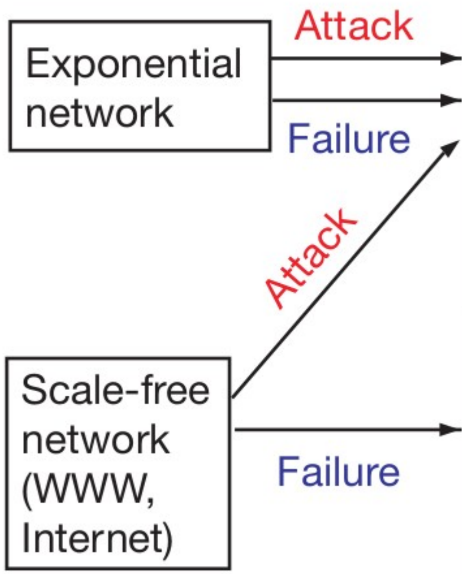
S: Fraction of nodes in largest cluster

<s>: average size of isolated clusters



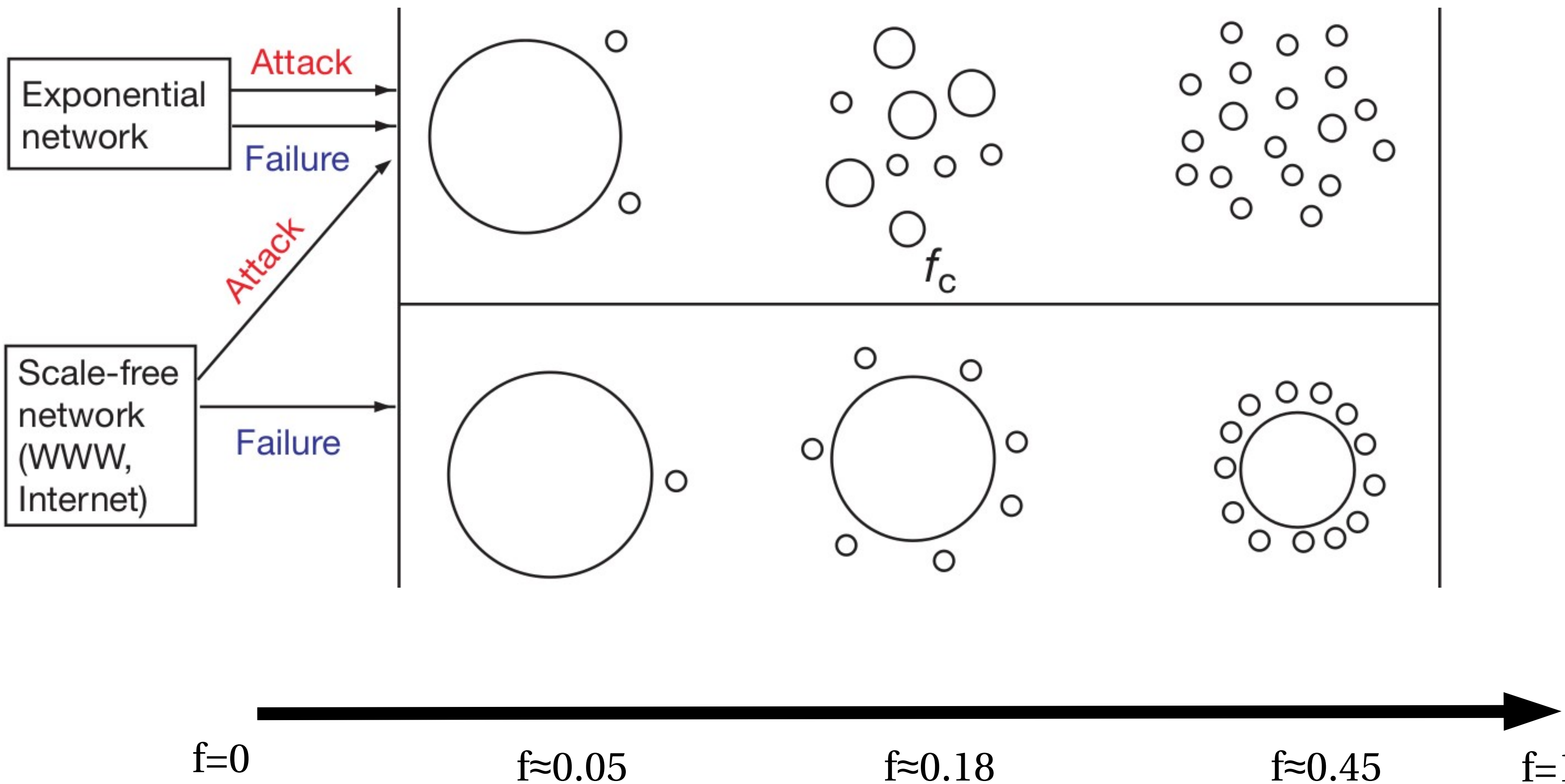
Albert, Réka, et al. **"Error and attack tolerance of complex networks."** nature 406, no. 6794 (2000): 378-382.

Network Response to Attacks and Failures



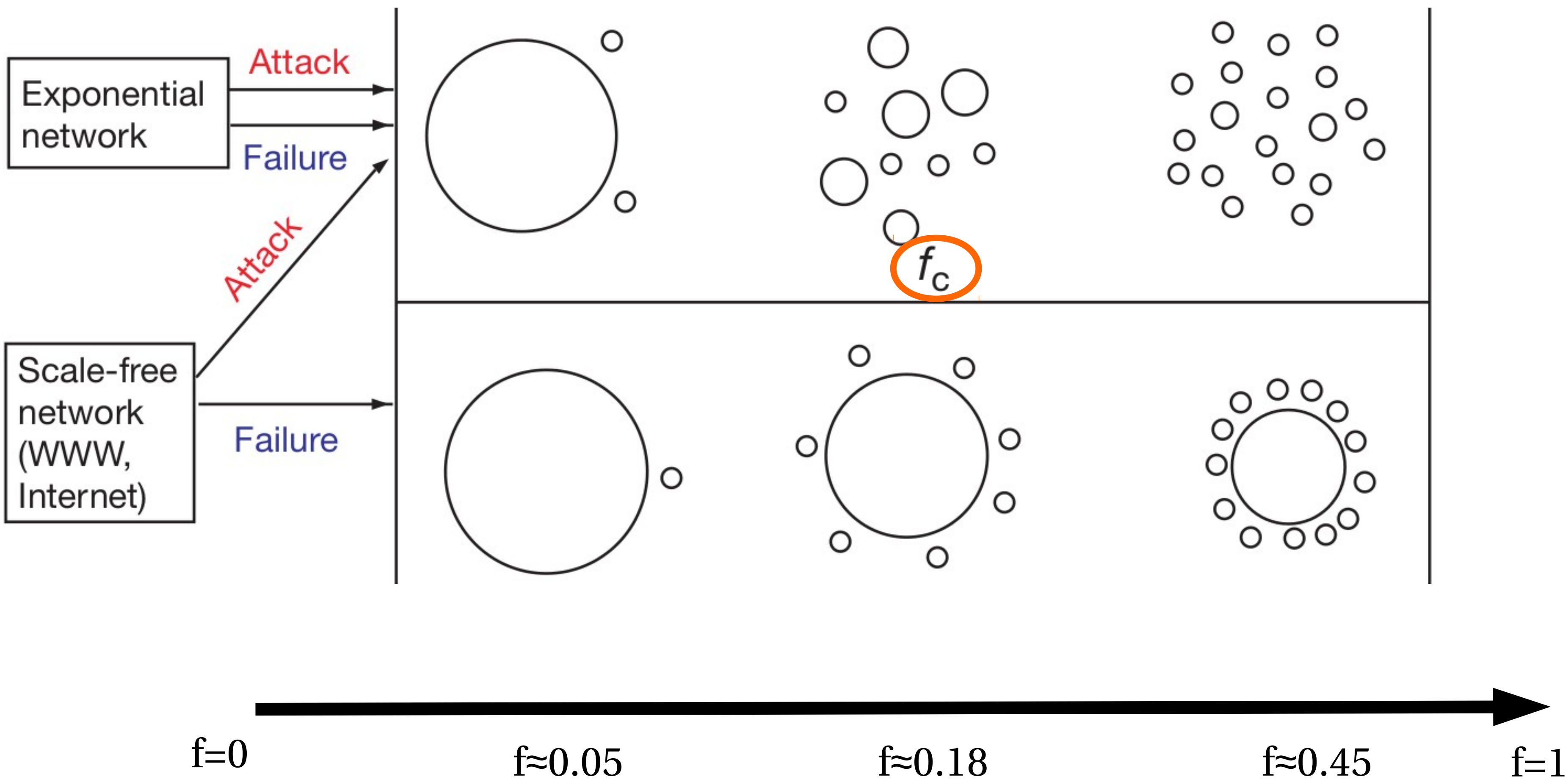
Albert, Réka, et al. **"Error and attack tolerance of complex networks."** nature 406, no. 6794 (2000): 378-382.

Network Response to Attacks and Failures



Albert, Réka, et al. **"Error and attack tolerance of complex networks."** nature 406, no. 6794 (2000): 378-382.

Network Response to Attacks and Failures



Albert, Réka, et al. **"Error and attack tolerance of complex networks."** nature 406, no. 6794 (2000): 378-382.

Critical Threshold

(random node failures)

Cohen, Reuven et al. **"Resilience of the Internet to random breakdowns."** Physical review letters 85, no. 21 (2000): 4626.

Critical Threshold

(random node failures)

$$\begin{cases} \gamma : \text{ exponent of power-law} \\ m : \text{ smallest degree} \\ N : \text{ number of nodes in the graph} \\ K : \text{ largest degree },\ K \approx mN^{\frac{1}{\gamma-1}} \end{cases}$$

Cohen, Reuven et al. **"Resilience of the Internet to random breakdowns."** Physical review letters 85, no. 21 (2000): 4626.

Critical Threshold

(random node failures)

$$f_c = 1 - \frac{1}{\beta - 1} \quad \begin{cases} \gamma : \text{ exponent of power-law} \\ m : \text{ smallest degree} \\ N : \text{ number of nodes in the graph} \\ K : \text{ largest degree }, K \approx mN^{\frac{1}{\gamma-1}} \end{cases}$$

Cohen, Reuven et al. **"Resilience of the Internet to random breakdowns."** Physical review letters 85, no. 21 (2000): 4626.

Critical Threshold

(random node failures)

$$f_c = 1 - \frac{1}{\beta - 1} \quad \begin{cases} \gamma : \text{ exponent of power-law} \\ m : \text{ smallest degree} \\ N : \text{ number of nodes in the graph} \\ K : \text{ largest degree },\ K \approx mN^{\frac{1}{\gamma-1}} \end{cases}$$

where

$$\beta = \frac{|2-\gamma|}{|3-\gamma|} \times \begin{cases} m & \text{if } \gamma > 3 \\ m^{\gamma-2}K^{3-\gamma} & \text{if } 2 < \gamma < 3 \\ K & \text{if } 1 < \gamma < 2 \end{cases}$$

Cohen, Reuven et al. **"Resilience of the Internet to random breakdowns."** Physical review letters 85, no. 21 (2000): 4626.

Critical Threshold

(random node failures)

$$f_c = 1 - \frac{1}{\beta - 1} \quad \begin{cases} \gamma : \text{ exponent of power-law} \\ m : \text{ smallest degree} \\ N : \text{ number of nodes in the graph} \\ K : \text{ largest degree }, \ K \approx mN^{\frac{1}{\gamma-1}} \end{cases}$$

where

$$\beta = \frac{|2-\gamma|}{|3-\gamma|} \times \begin{cases} m & \text{if } \gamma > 3 \\ m^{\gamma-2}K^{3-\gamma} & \text{if } 2 < \gamma < 3 \\ K & \text{if } 1 < \gamma < 2 \end{cases}$$

for $2 < \gamma < 3$

Cohen, Reuven et al. **"Resilience of the Internet to random breakdowns."** Physical review letters 85, no. 21 (2000): 4626.

Critical Threshold

(random node failures)

$$f_c = 1 - \frac{1}{\beta - 1} \quad \begin{cases} \gamma : \text{ exponent of power-law} \\ m : \text{ smallest degree} \\ N : \text{ number of nodes in the graph} \\ K : \text{ largest degree }, \; K \approx mN^{\frac{1}{\gamma-1}} \end{cases}$$

where

$$\beta = \frac{|2-\gamma|}{|3-\gamma|} \times \begin{cases} m & \text{if } \gamma > 3 \\ m^{\gamma-2}K^{3-\gamma} & \text{if } 2 < \gamma < 3 \\ K & \text{if } 1 < \gamma < 2 \end{cases}$$

for $2 < \gamma < 3$

$$f_c = 1 + \left(1 - m^{(\gamma-2)}K^{(3-\gamma)}\frac{\gamma-2}{3-\gamma}\right)^{-1}$$

Cohen, Reuven et al. **"Resilience of the Internet to random breakdowns."** Physical review letters 85, no. 21 (2000): 4626.

Critical Threshold

(random node failures)

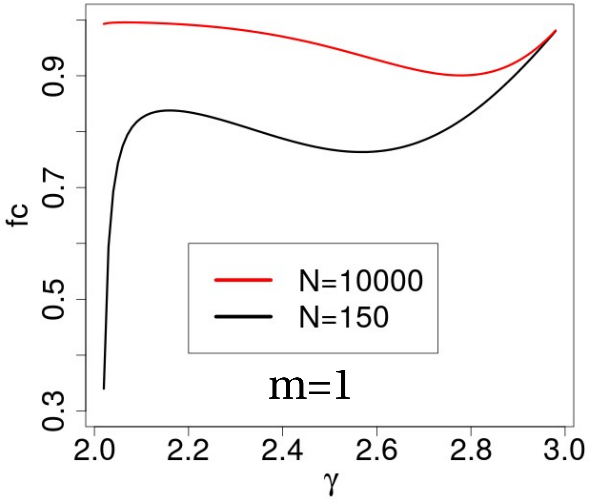
$$f_c = 1 - \frac{1}{\beta - 1} \quad \begin{cases} \gamma : \text{ exponent of power-law} \\ m : \text{ smallest degree} \\ N : \text{ number of nodes in the graph} \\ K : \text{ largest degree }, \; K \approx mN^{\frac{1}{\gamma-1}} \end{cases}$$

where

$$\beta = \frac{|2-\gamma|}{|3-\gamma|} \times \begin{cases} m & \text{if } \gamma > 3 \\ m^{\gamma-2}K^{3-\gamma} & \text{if } 2 < \gamma < 3 \\ K & \text{if } 1 < \gamma < 2 \end{cases}$$

for $2 < \gamma < 3$

$$f_c = 1 + \left(1 - m^{(\gamma-2)}K^{(3-\gamma)}\frac{\gamma-2}{3-\gamma}\right)^{-1}$$

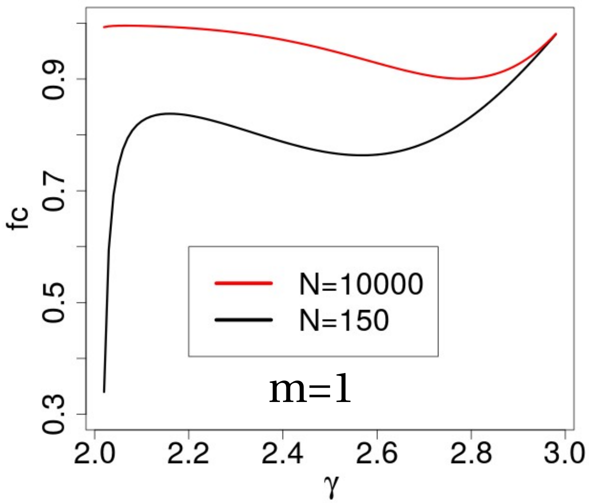


Cohen, Reuven et al. **"Resilience of the Internet to random breakdowns."** Physical review letters 85, no. 21 (2000): 4626.

Critical Threshold

(random node failures)

$$f_c = 1 - \frac{1}{\beta - 1} \quad \begin{cases} \gamma : \text{ exponent of power-law} \\ m : \text{ smallest degree} \\ N : \text{ number of nodes in the graph} \\ K : \text{ largest degree }, \; K \approx mN^{\frac{1}{\gamma-1}} \end{cases}$$



where

$$\beta = \frac{|2-\gamma|}{|3-\gamma|} \times \begin{cases} m & \text{if } \gamma > 3 \\ m^{\gamma-2}K^{3-\gamma} & \text{if } 2 < \gamma < 3 \\ K & \text{if } 1 < \gamma < 2 \end{cases}$$

for $2 < \gamma < 3$

$$f_c = 1 + \left(1 - m^{(\gamma-2)}K^{(3-\gamma)}\frac{\gamma-2}{3-\gamma}\right)^{-1}$$

Cohen's technique can be extended to errors

(No closed form for f_c for errors)

Cohen, Reuven et al. "**Resilience of the Internet to random breakdowns.**" Physical review letters 85, no. 21 (2000): 4626.

Summary on Attack and Error Tolerance of Complex Networks

Scale-free networks resilient to random failures but vulnerable to targetted attacks