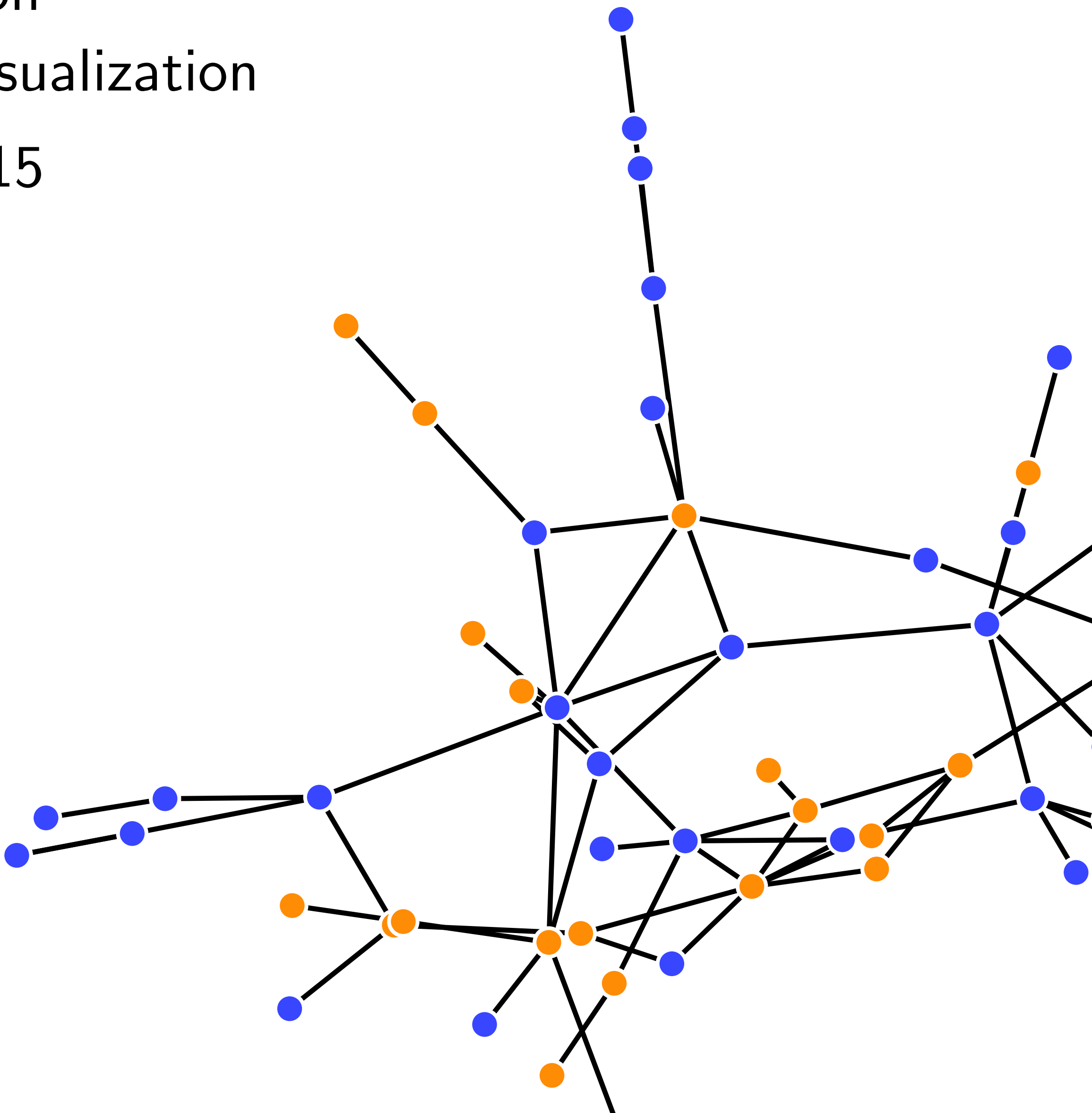
A Tale of two communities
Assessing Homophily in Node-Link Diagrams

23rd International Symposium on
Graph-Drawing and Network Visualization

Los Angeles, September 26, 2015



Wouter Meulemans
City University London

André Schulz
FernUniversität in Hagen

Homophily

Homophily

- **homophily** is a concept in social network analysis

Homophily

- **homophily** is a concept in social network analysis
- more likely that two individuals with a common charactristic form a link $\rightarrow$ homophily

Homophily

- **homophily** is a concept in social network analysis
- more likely that two individuals with a common charactristic form a link $\rightarrow$ homophily
  (example: same-gender links are more likely in a friendship-networks)

Homophily

- **homophily** is a concept in social network analysis
- more likely that two individuals with a common charactristic form a link $\rightarrow$ homophily

  (example: same-gender links are more likely in a friendship-networks)
- **reason 1 for homophily:** “Birds of feather flock together”

  (social selection)

Homophily

- **homophily** is a concept in social network analysis
- more likely that two individuals with a common charactristic form a link $\rightarrow$ homophily

  (example: same-gender links are more likely in a friendship-networks)
- **reason 1 for homophily:** "Birds of feather flock together"

  (social selection)
- **reason 2 for homophily:** we form characteristics similar to our friends **(social influence)**

Homophily

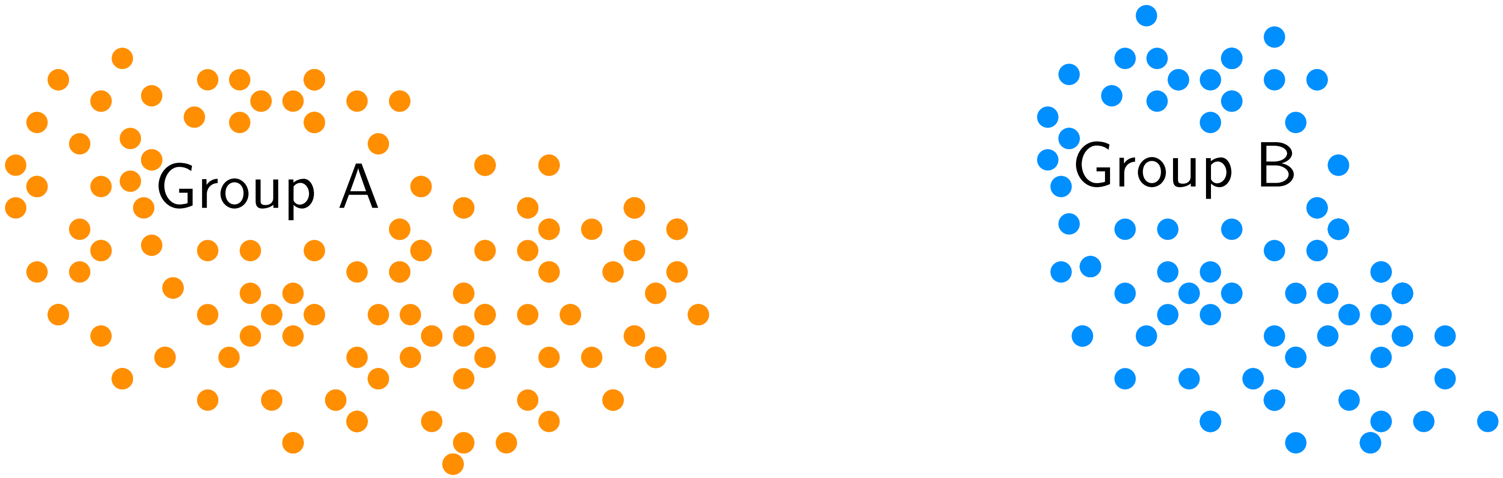
- **homophily** is a concept in social network analysis
- more likely that two individuals with a common charactristic form a link → homophily
  (example: same-gender links are more likely in a friendship-networks)
- **reason 1 for homophily:** "Birds of feather flock together"
  (social selection)
- **reason 2 for homophily:** we form characteristics similar to our friends **(social influence)**
- also effects opposite to homophily can occur (heterophily)

Homophily

- **homophily** is a concept in social network analysis
- more likely that two individuals with a common charactristic form a link $\rightarrow$ homophily
  (example: same-gender links are more likely in a friendship-networks)
- **reason 1 for homophily:** "Birds of feather flock together"
  (social selection)
- **reason 2 for homophily:** we form characteristics similar to our friends **(social influence)**
- also effects opposite to homophily can occur (heterophily)
- homophily is not restricted to *social networks*
  (Question: groups = clusters?)

Formalizing Homophily

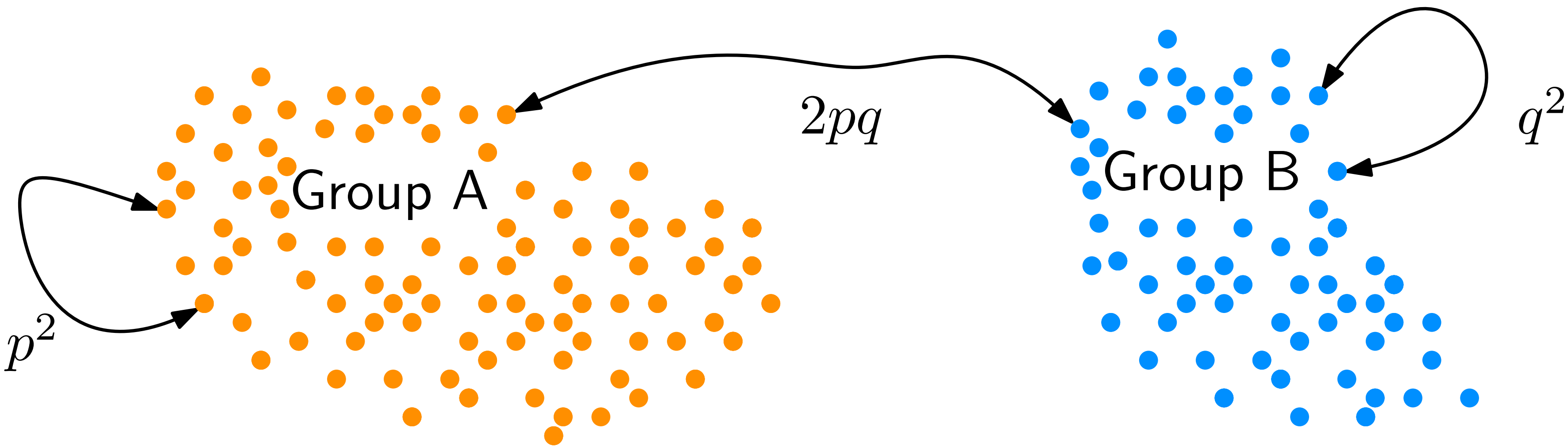
Formalizing Homophily



fraction p of the individuals

fraction q of the individuals

Formalizing Homophily



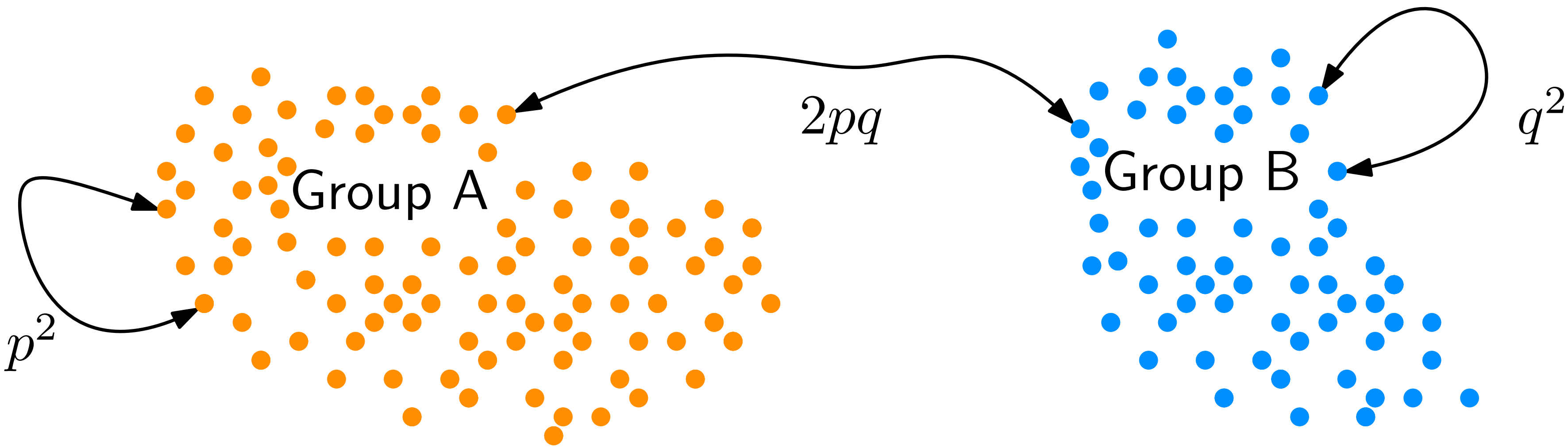
fraction p of the individuals

fraction q of the individuals

A random link is

- with probability p^2: A $\leftrightarrow$ A
- with probability q^2: B $\leftrightarrow$ B
- with probability $2pq$: A $\leftrightarrow$ B

Formalizing Homophily



fraction p of the individuals

fraction q of the individuals

A random link is

- with probability p^2: A $\leftrightarrow$ A
- with probability q^2: B $\leftrightarrow$ B
- with probability $2pq$: A $\leftrightarrow$ B

Homophily Test

If the fraction of the between-group links is significantly smaller than $2pq$ we have homophily.

Degree of Homophily

- we want to measure the **degree of homophily** in a network

Degree of Homophily

- we want to measure the **degree of homophily** in a network

Important Cases

① only cross-group links (heterophily)

② $2pq$ cross-group links (balanced)

③ no cross-group links (homophily)

Degree of Homophily

- we want to measure the **degree of homophily** in a network

Important Cases

①	only cross-group links (heterophily)	**0**
②	$2pq$ cross-group links (balanced)	**1/2**
③	no cross-group links (homophily)	**1**

Degree of Homophily

Degree of Homophily

- we want to measure the **degree of homophily** in a network

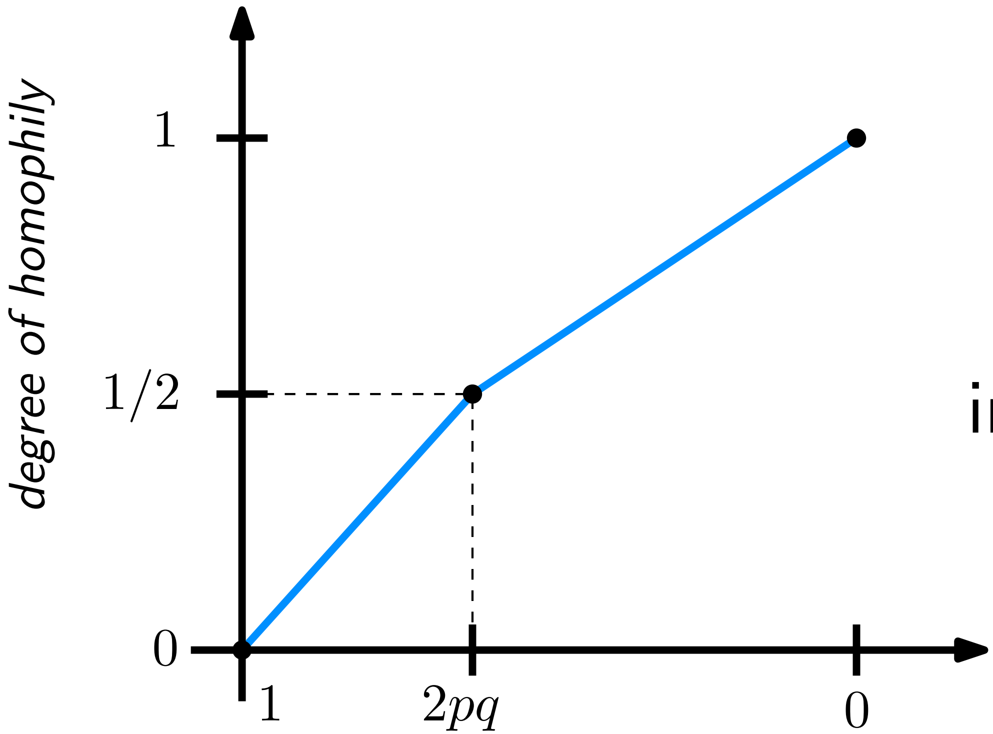
Important Cases

① only cross-group links (heterophily) **0**

② $2pq$ cross-group links (balanced) **1/2**

③ no cross-group links (homophily) **1**

Degree of Homophily



interpolate all other values linearly

Research Questions

Research Questions

Can an observer assess homophily
in a node-link diagram?

Research Questions

Can an observer assess homophily
in a node-link diagram?

Subquestions:

- Which node-link diagram layout is best suitable for detecting homophily?

Research Questions

Can an observer assess homophily
in a node-link diagram?

Subquestions:

- Which node-link diagram layout is best suitable for detecting homophily?
- Is there a tendency for overestimation or underestimation?

Research Questions

Can an observer assess homophily
in a node-link diagram?

Subquestions:

- Which node-link diagram layout is best suitable for detecting homophily?
- Is there a tendency for overestimation or underestimation?
- Are there general design principles to improve homophily detection?

Research Questions

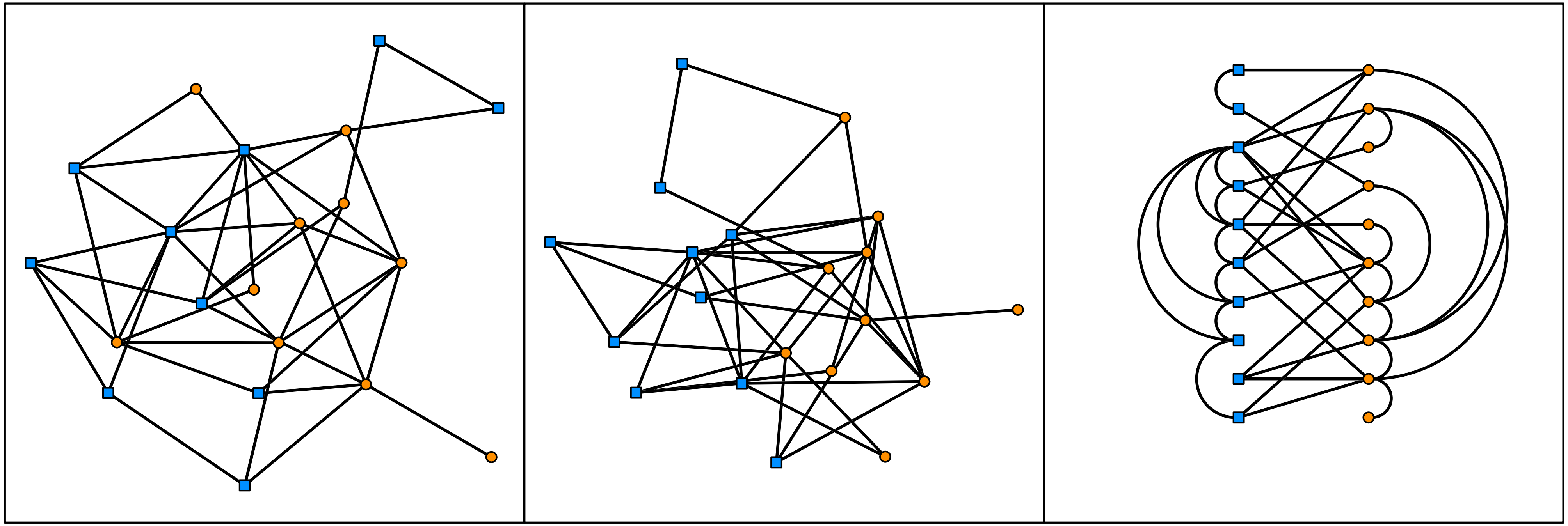
Can an observer assess homophily
in a node-link diagram?

Subquestions:

- Which node-link diagram layout is best suitable for detecting homophily?
- Is there a tendency for overestimation or underestimation?
- Are there general design principles to improve homophily detection?

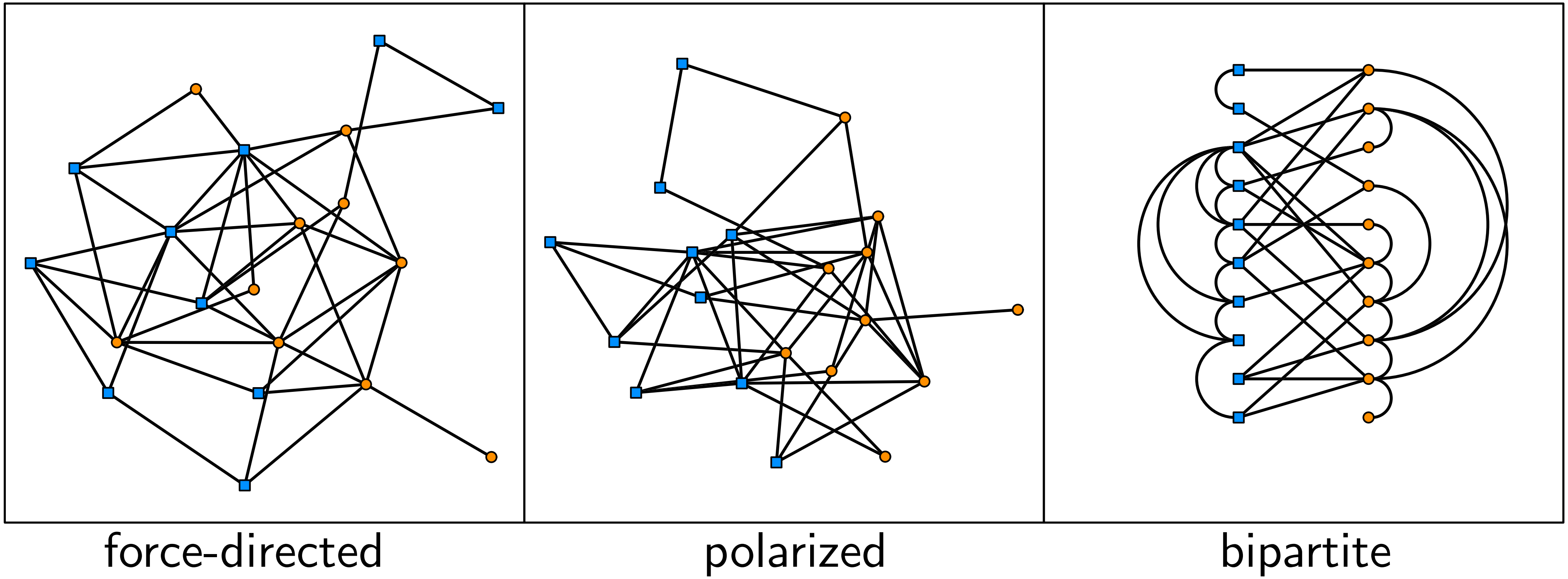
ⓘ **We only consider node-link diagrams and the "two-groups-scenario"**

Layouts



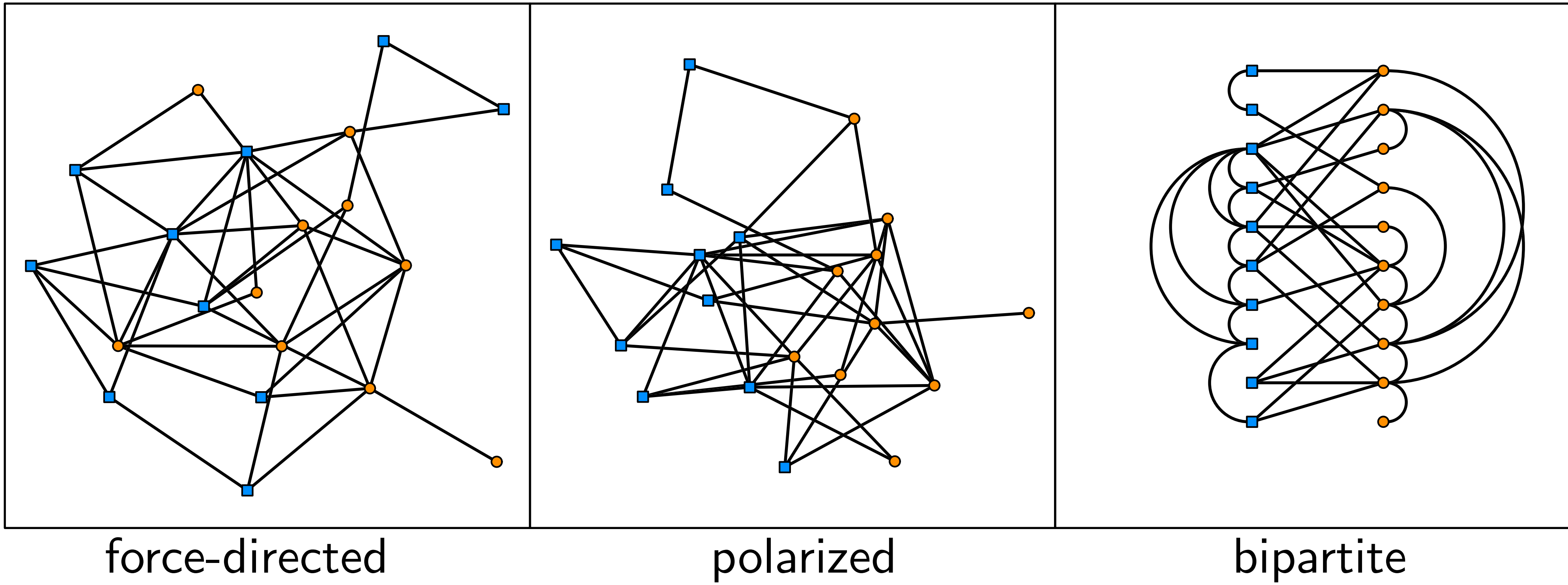
force-directed polarized bipartite

Layouts



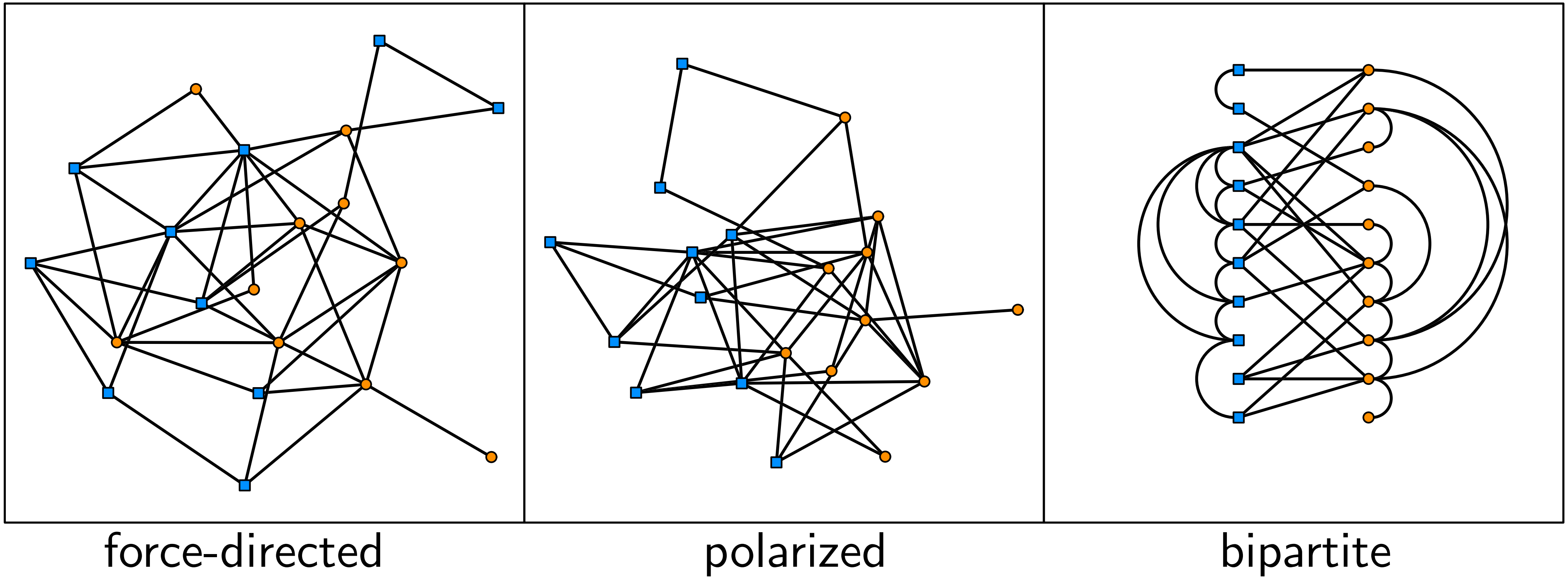
- layout based on the Fruchtermann–Reingold Algorithm
- implementation taken from the `d3.js` library

Layouts



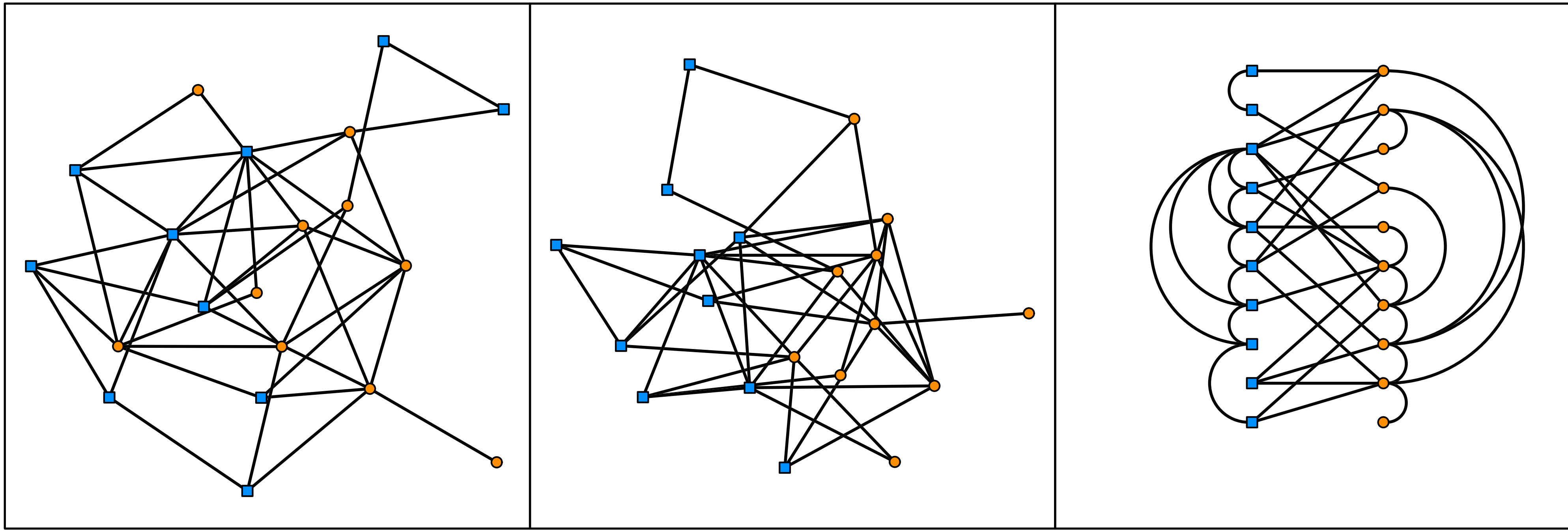
- modification of the force-directed layout
- additional forces pull blue vertices to the left and red vertices to the right

Layouts



- groups are placed on opposing vertical lines
- barycentric layout + sifting to remove crossings
- different shapes for cross-group/within-group edges

Layouts



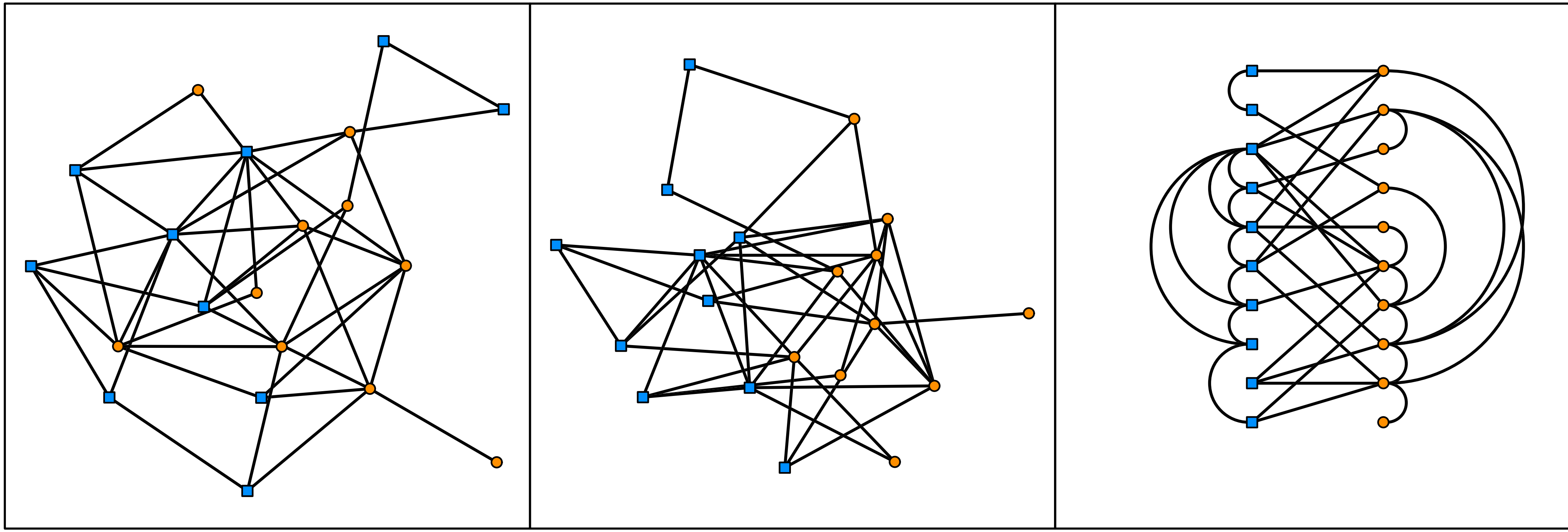
force-directed

polarized

bipartite

group separation

Layouts



force-directed

polarized

bipartite

group separation

homophily detection easier?

other tasks more difficult?

Hypothesis

Hypothesis

H1 For Homophily assessment we have

force-directed $<$ **polarized** $<$ **bipartite**

$x < y$ means y is better than x

Hypothesis

H1 For Homophily assessment we have

$$\textbf{force-directed} < \textbf{polarized} < \textbf{bipartite}$$

$x < y$ means y is better than x

H2 For Homophily assesment we have

$$\textbf{unbalanced} < \textbf{balanced}$$

Hypothesis

H1 For Homophily assessment we have

$$\textbf{force-directed} < \textbf{polarized} < \textbf{bipartite}$$

$x < y$ means y is better than x

H2 For Homophily assesment we have

$$\textbf{unbalanced} < \textbf{balanced}$$

H3 For shortest path queries we have

$$\textbf{force-directed} > \textbf{polarized} > \textbf{bipartite}$$

User Study Design

User Study Design

mixed design (too much trials otherwise)

User Study Design

mixed design (too much trials otherwise)

between subject

- 3 graph sizes (20-28 nodes, 20-40 edges)

User Study Design

mixed design (too much trials otherwise)

between subject

- 3 graph sizes (20-28 nodes, 20-40 edges)

within subjects

- 3 layouts
- balanced (50:50) and unbalanced (25:75)
- 5 degree of homophily levels (only 3 for unbalanced)
- 2 tasks (homophily / length of shortest path)

User Study Design

mixed design (too much trials otherwise)

between subject

- 3 graph sizes (20-28 nodes, 20-40 edges)

within subjects

- 3 layouts
- balanced (50:50) and unbalanced (25:75)
- 5 degree of homophily levels (only 3 for unbalanced)
- 2 tasks (homophily / length of shortest path)

demo of the user study

`http://tutte.fernuni-hagen.de/~schulza`

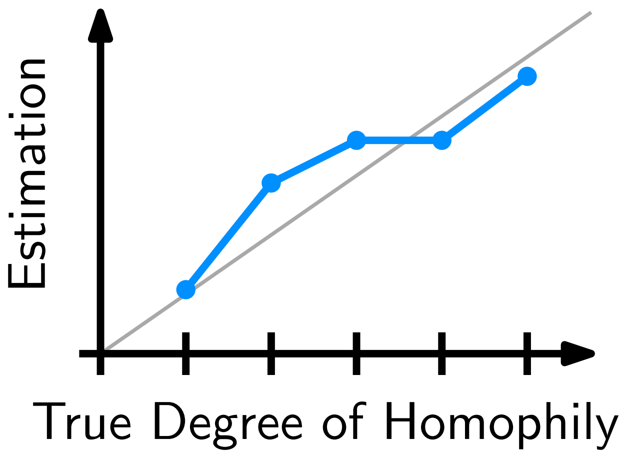
Evaluating Results

Evaluating Results

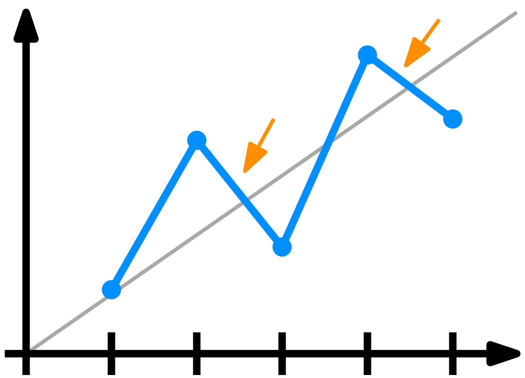
- Users have an internal “scale” for the degree of homophily

Evaluating Results

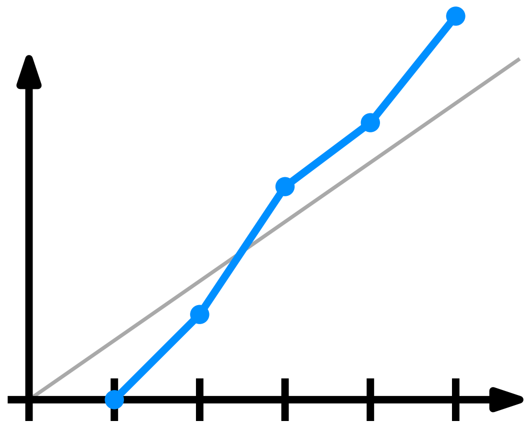
- Users have an internal “scale” for the degree of homophily



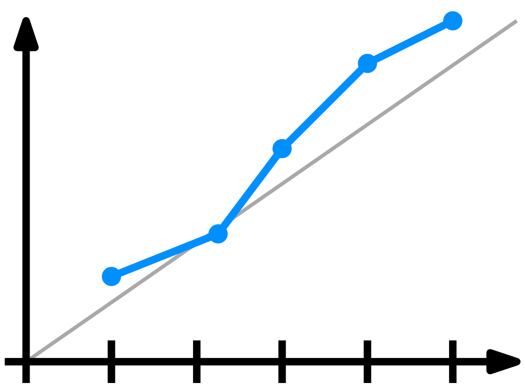
good estimation



bad estimation

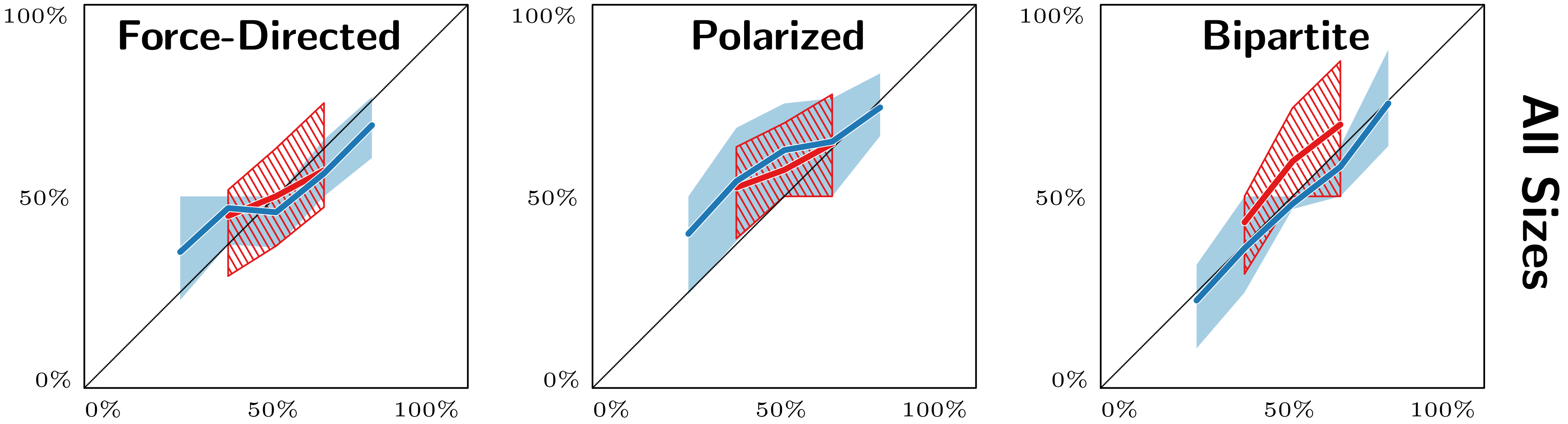


good estimation (different personal scale)

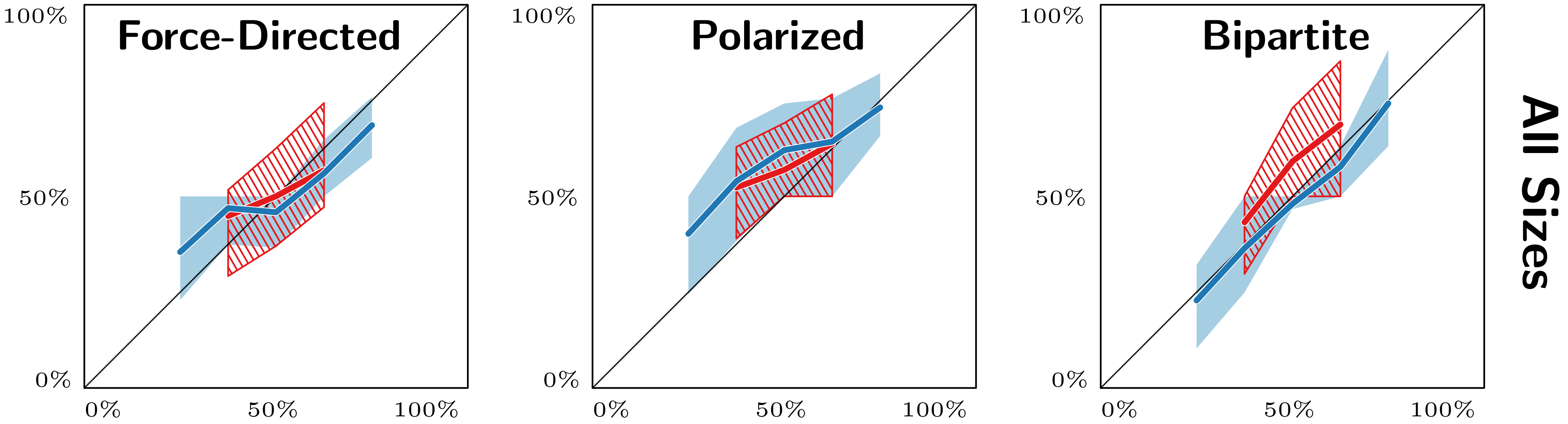


okay estimation but overestimated

Homophily Results

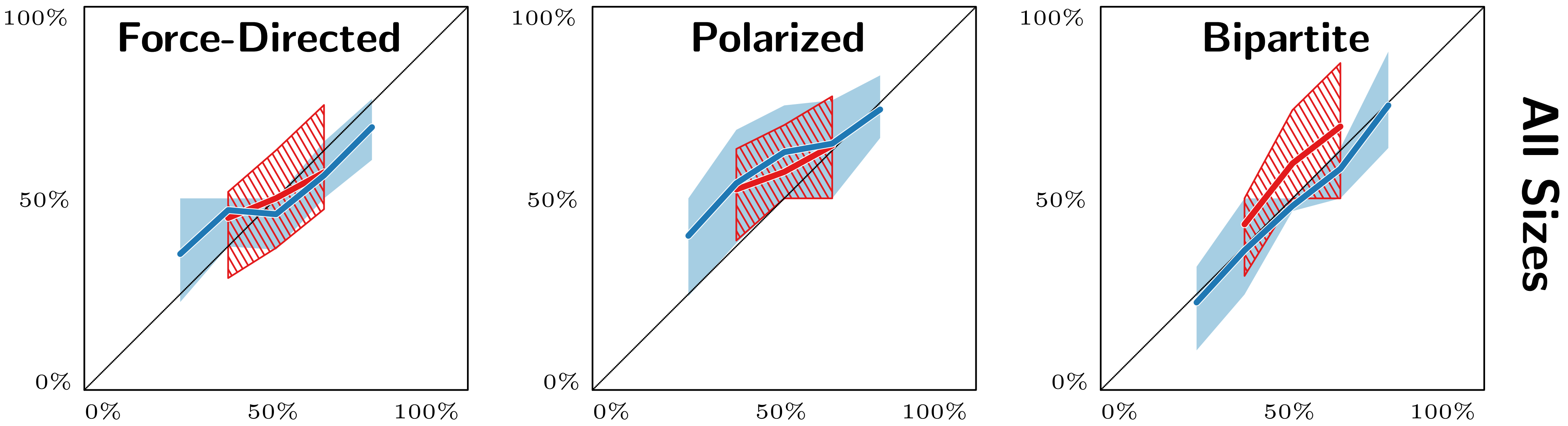


Homophily Results



- polarized < bipartite, force-directed

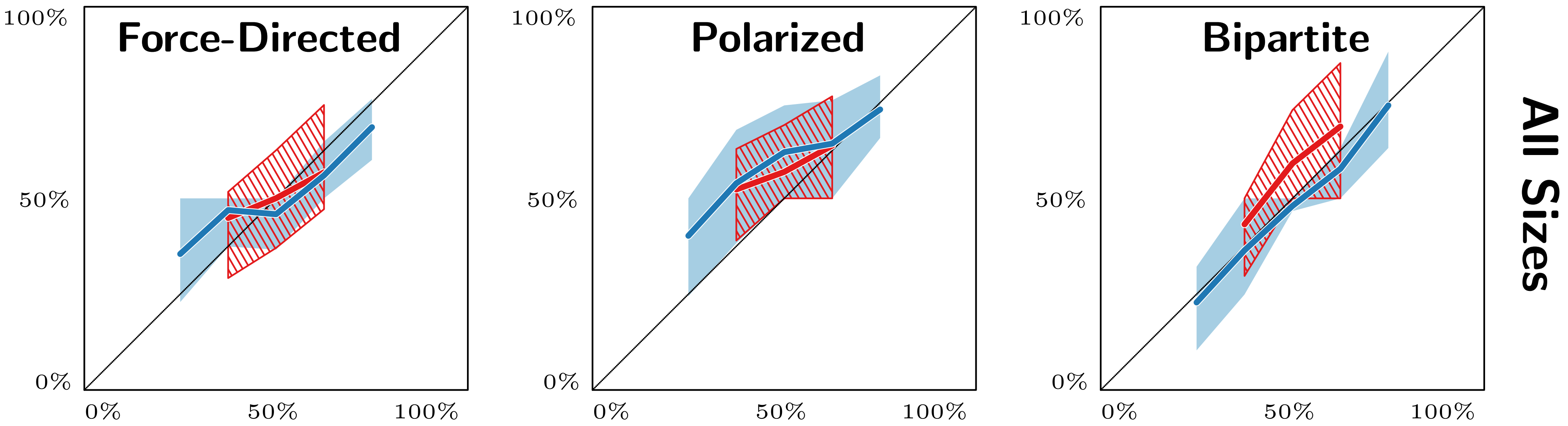
Homophily Results



- polarized < bipartite, force-directed

statistical evidence

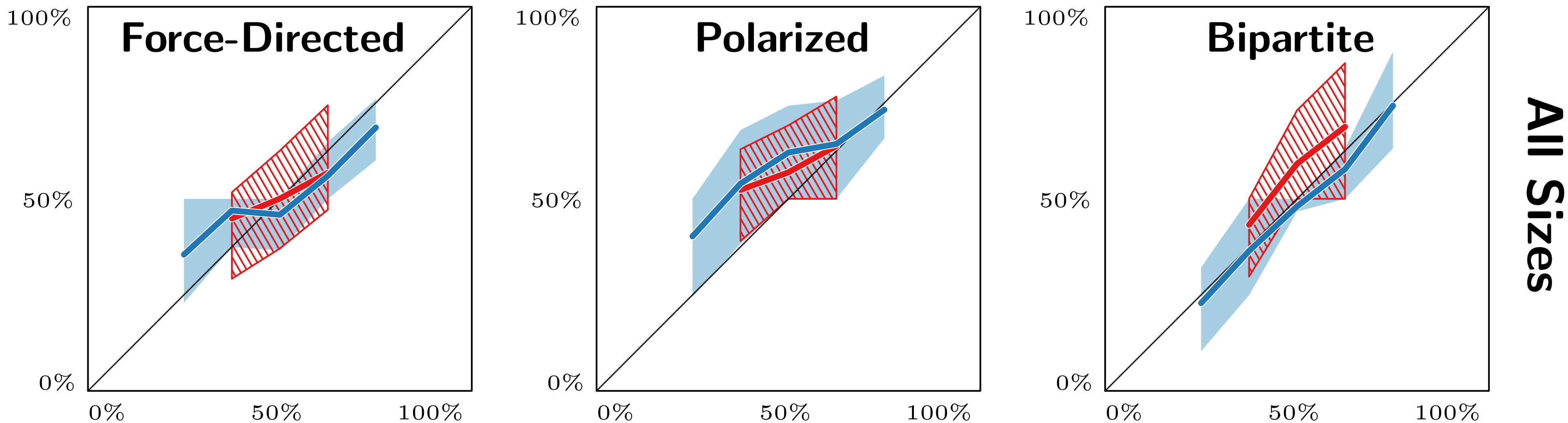
Homophily Results



- polarized < bipartite, force-directed
- no difference between force-direced and bipartite

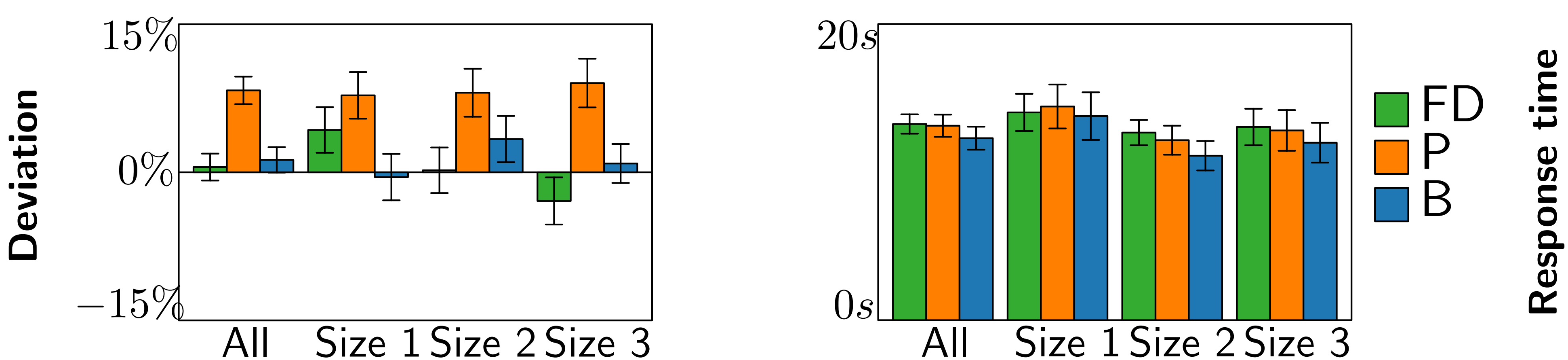
statistical evidence

Homophily Results

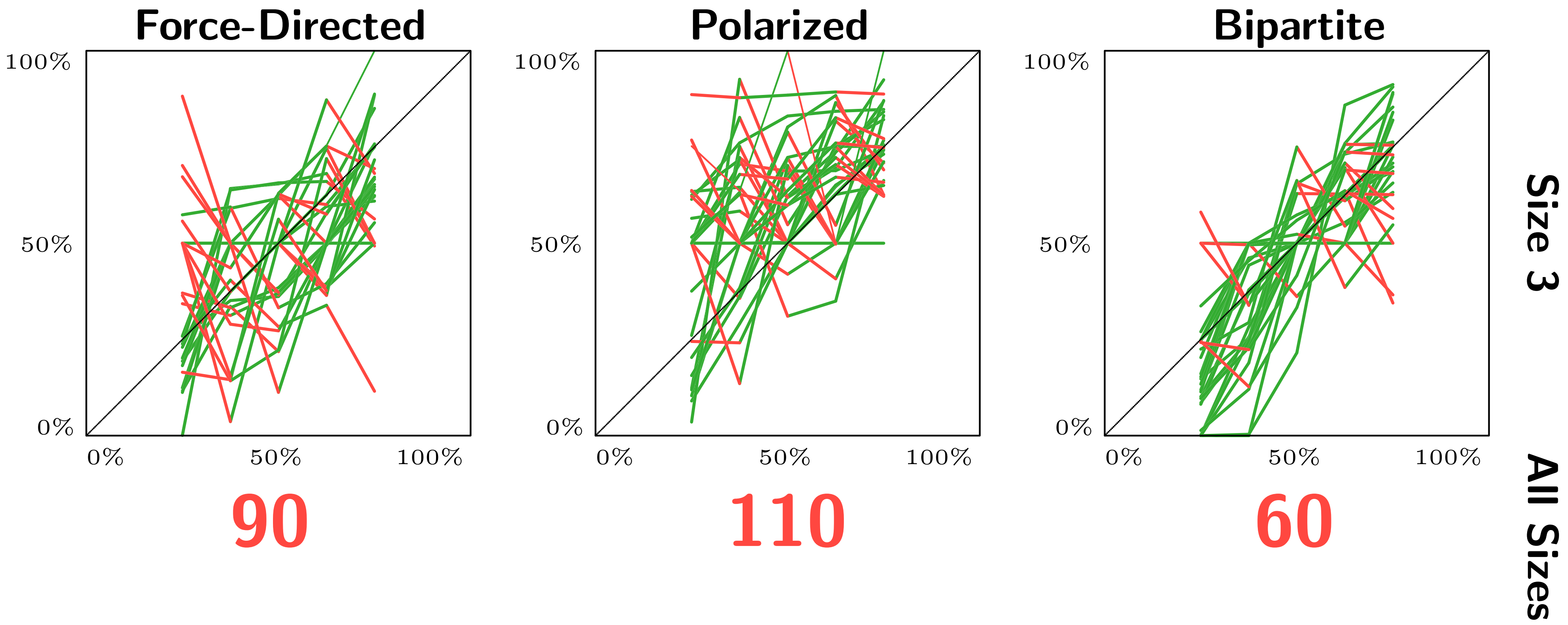


- polarized < bipartite, force-directed
- no difference between force-direced and bipartite

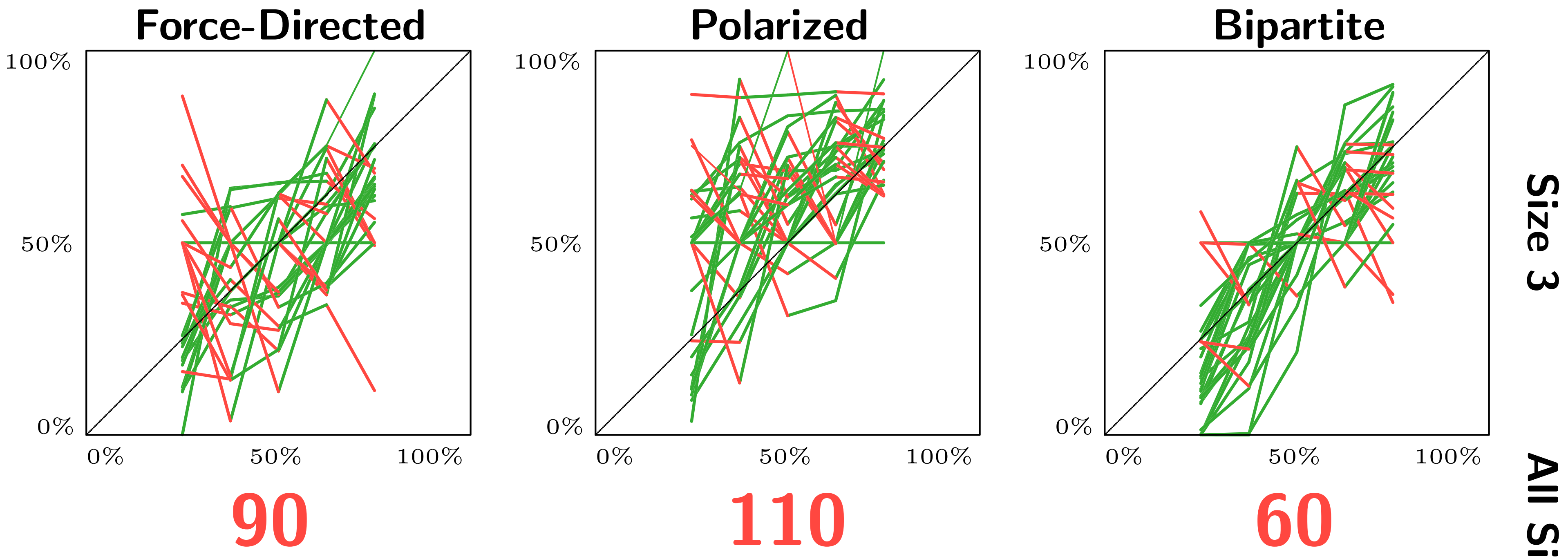
statistical evidence



Homophily Results - Internal Consistency

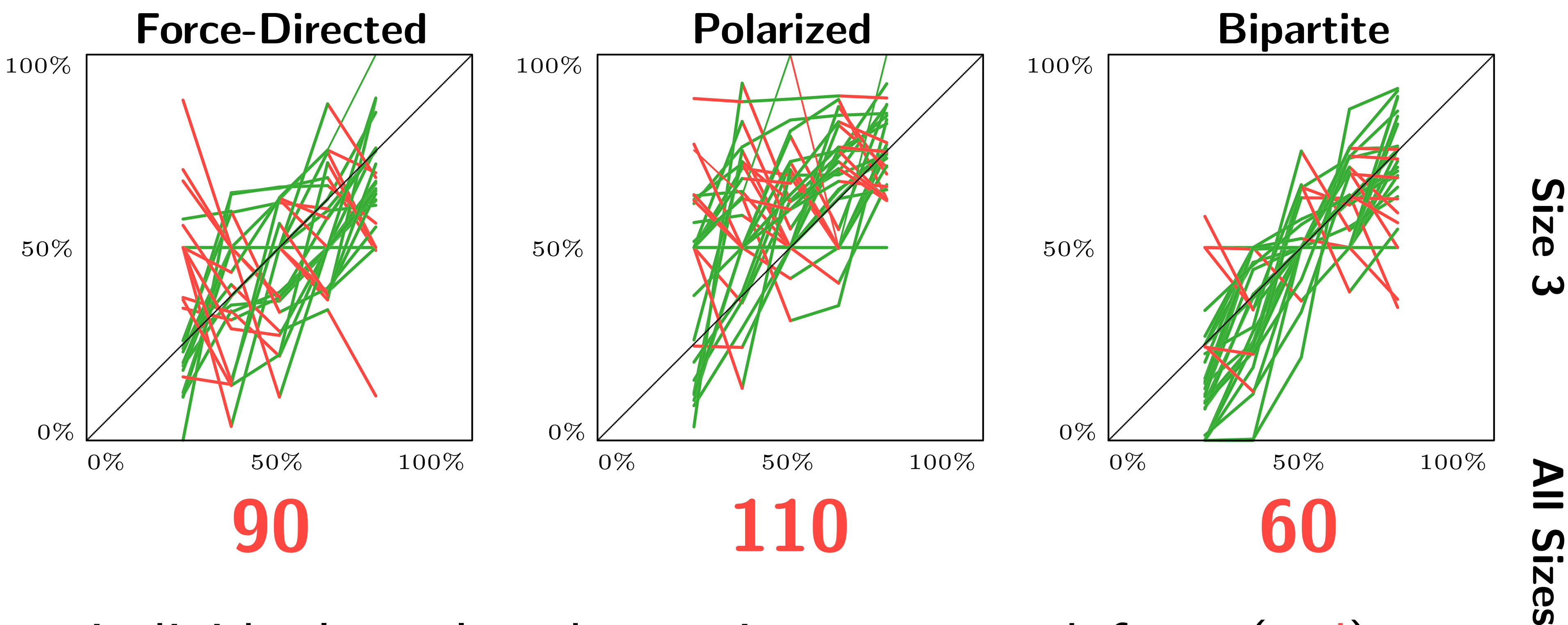


Homophily Results - Internal Consistency



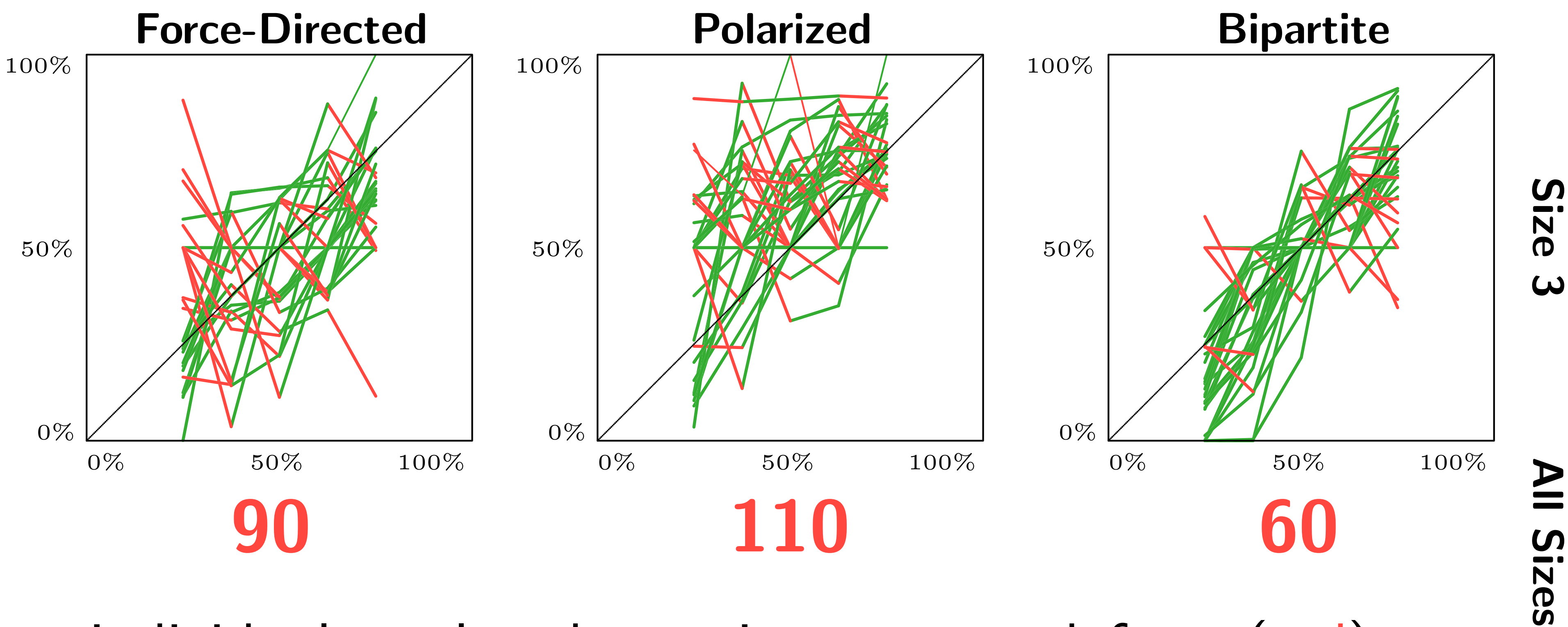
- individual results, decreasing parts = defects (red)

Homophily Results - Internal Consistency



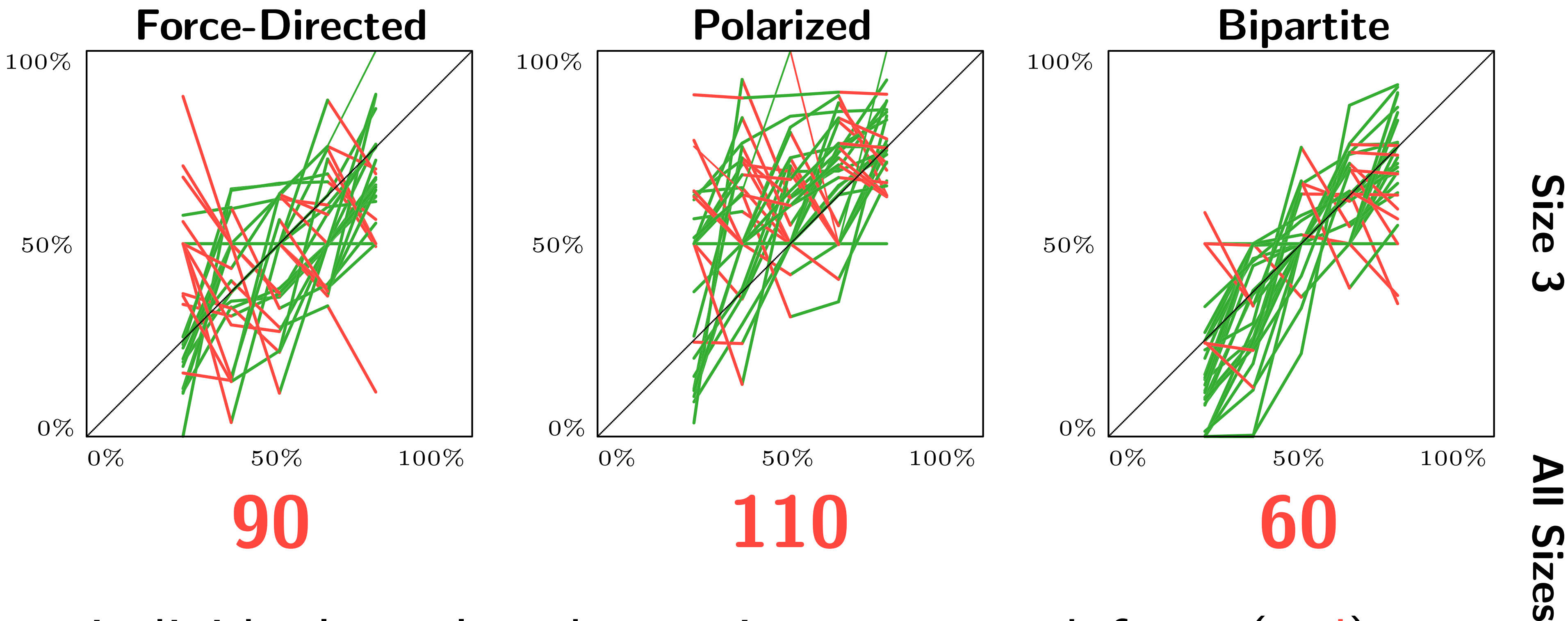
- individual results, decreasing parts = defects (red)
- many inconsistencies (not clear from the aggregated data)

Homophily Results - Internal Consistency



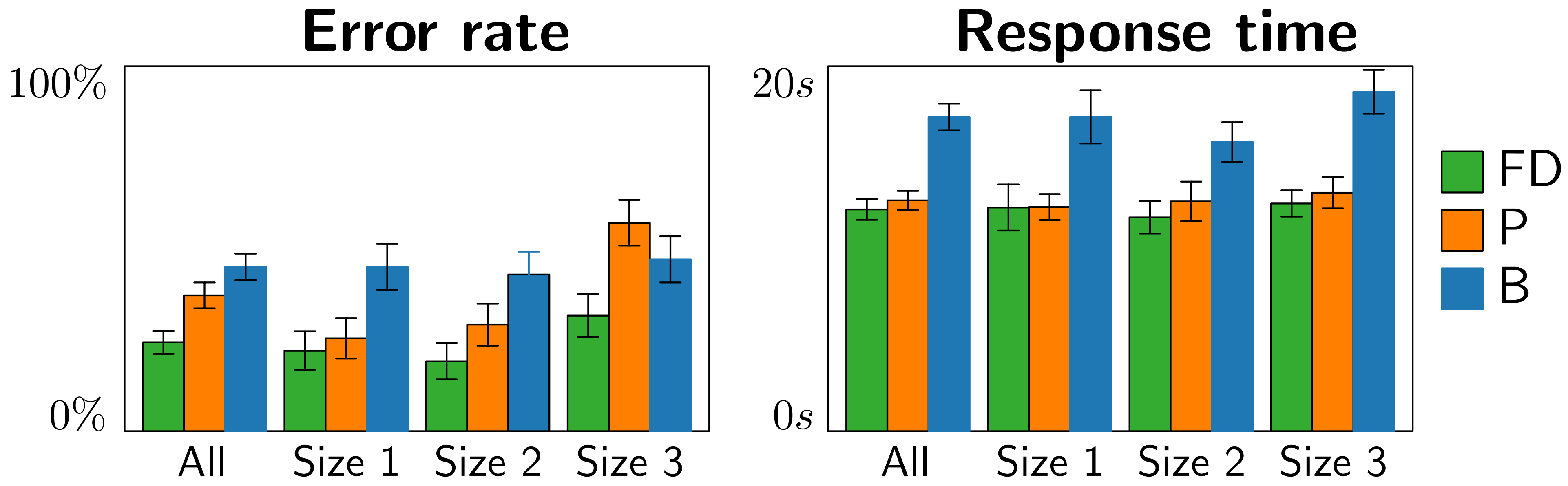
- individual results, decreasing parts = defects (red)
- many inconsistencies (not clear from the aggregated data)
- evidence that **bipartite** > **force-directed**

Homophily Results - Internal Consistency

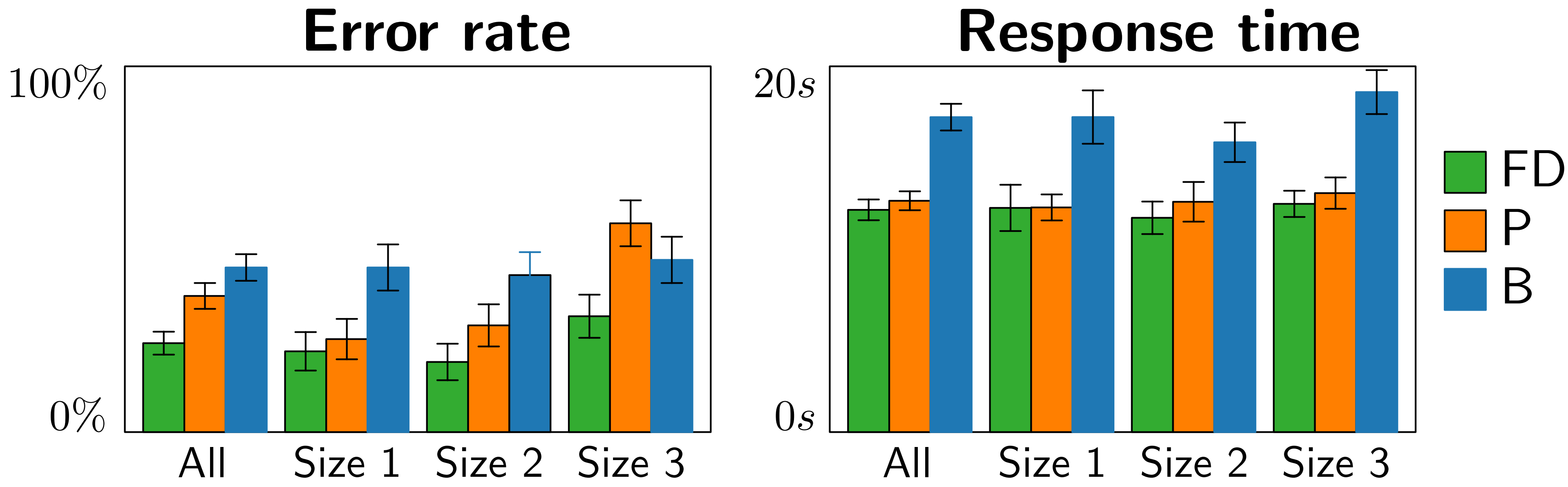


- individual results, decreasing parts = defects (red)
- many inconsistencies (not clear from the aggregated data)
- evidence that **bipartite** > **force-directed**
- tendency to overestimate in the polarized layout

Shortest Path Results



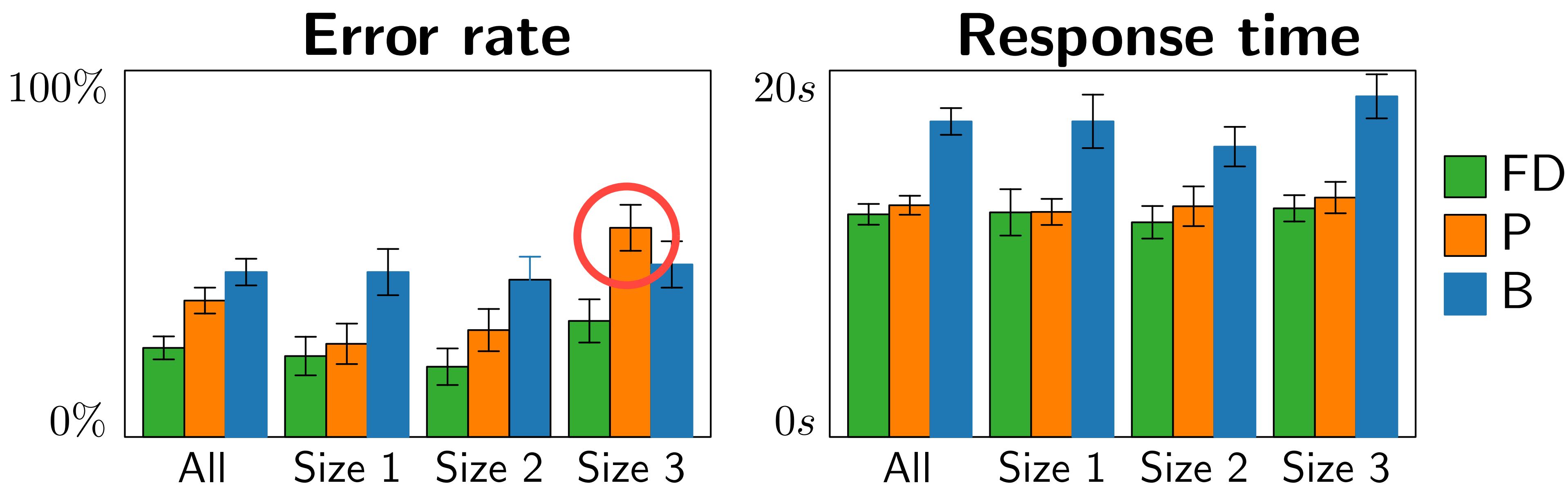
Shortest Path Results



forced-directed better than **polarized** better than **bipartite**

(again supported by statistical evidence)

Shortest Path Results

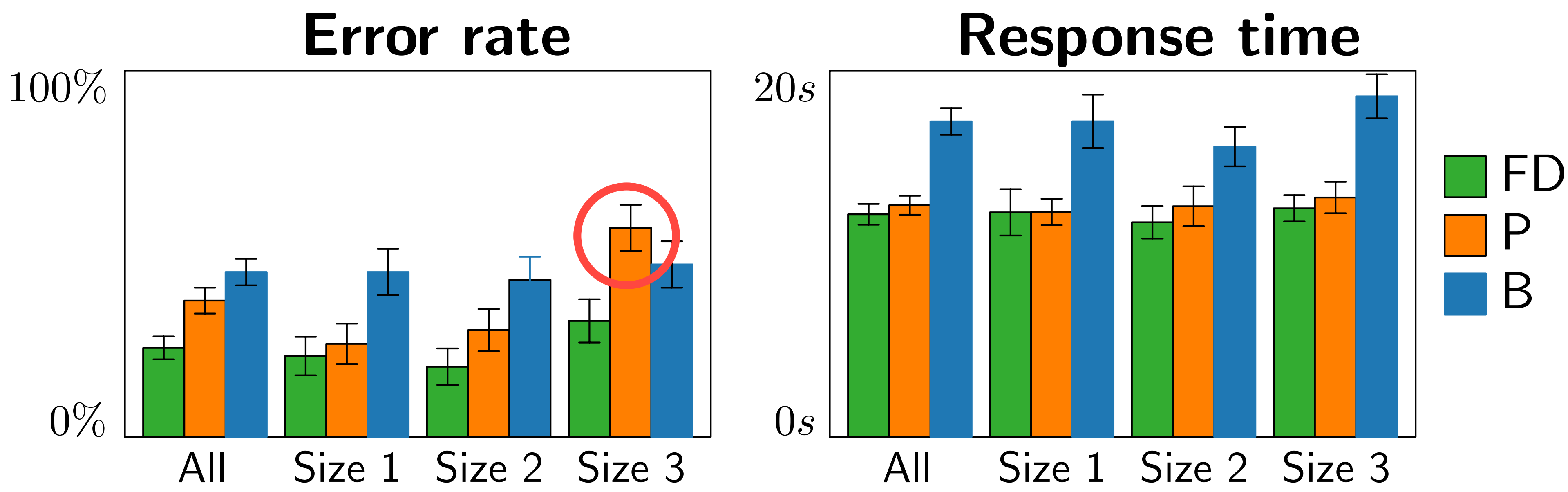


forced-directed better than **polarized** better than **bipartite**

(again supported by statistical evidence)

- there was one problematic instance in size group 3 for the bipartite layout, caused by collinearities in the layout

Shortest Path Results



forced-directed better than **polarized** better than **bipartite**

(again supported by statistical evidence)

- there was one problematic instance in size group 3 for the bipartite layout, caused by collinearities in the layout
- size was not a big influence

Summary

H1 For Homophily assessment we have

$$\textbf{force-directed} < \textbf{polarized} < \textbf{bipartite}$$

Summary

H1 For Homophily assessment we have

force-directed $<$ **polarized** $<$ **bipartite**

- we can only partially accept **H1**:
  polarized $<$ **bipartite** **polarized** $<$ **force-directed**
- *internal consistency* data supports

  force-directed $<$ **bipartite**

Summary

H1 For Homophily assessment we have

$$\textbf{force-directed} < \textbf{polarized} < \textbf{bipartite}$$

- we can only partially accept **H1**:
  polarized $<$ **bipartite** **polarized** $<$ **force-directed**
- *internal consistency* data supports

$$\textbf{force-directed} < \textbf{bipartite}$$

H2 For Homophily assesment **unbalanced** $<$ **balanced**

H3 For shortest path queries we have

$$\textbf{force-directed} > \textbf{polarized} > \textbf{bipartite}$$

Summary

H1 For Homophily assessment we have

force-directed $<$ **polarized** $<$ **bipartite**

- we can only partially accept **H1**:
  polarized $<$ **bipartite** **polarized** $<$ **force-directed**
- *internal consistency* data supports

  force-directed $<$ **bipartite**

H2 For Homophily assesment **unbalanced** $<$ **balanced**

H3 For shortest path queries we have

force-directed $>$ **polarized** $>$ **bipartite**

- we can accept **H2** and **H3** based on our statistical analysis

Final thoughts

Final thoughts

- homophily is difficult to assess, but when averaging over a set of indivduals we get a good estimate

Final thoughts

- homophily is difficult to assess, but when averaging over a set of indivduals we get a good estimate
- the bipartite layout helped to assess homophily at the costs of more difficult path tracing

Final thoughts

- homophily is difficult to assess, but when averaging over a set of indivduals we get a good estimate
- the bipartite layout helped to assess homophily at the costs of more difficult path tracing
- node seperation, was not the primary reason for this, since the polarized layout was outperformed

Final thoughts

- homophily is difficult to assess, but when averaging over a set of indivduals we get a good estimate
- the bipartite layout helped to assess homophily at the costs of more difficult path tracing
- node seperation, was not the primary reason for this, since the polarized layout was outperformed
- the unbalanced case is harder

Final thoughts

- homophily is difficult to assess, but when averaging over a set of indivduals we get a good estimate
- the bipartite layout helped to assess homophily at the costs of more difficult path tracing
- node seperation, was not the primary reason for this, since the polarized layout was outperformed
- the unbalanced case is harder
- there is a tendency to overestimate in the polarized layout

Final thoughts

- homophily is difficult to assess, but when averaging over a set of indivduals we get a good estimate
- the bipartite layout helped to assess homophily at the costs of more difficult path tracing
- node seperation, was not the primary reason for this, since the polarized layout was outperformed
- the unbalanced case is harder
- there is a tendency to overestimate in the polarized layout

Thank you for your attention!