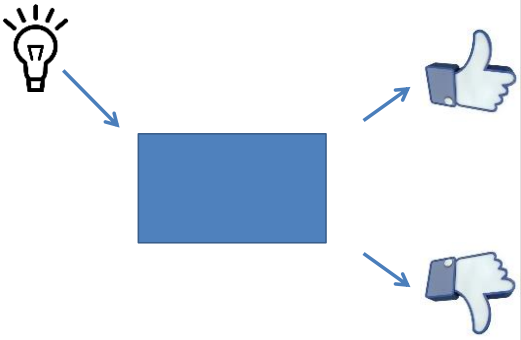
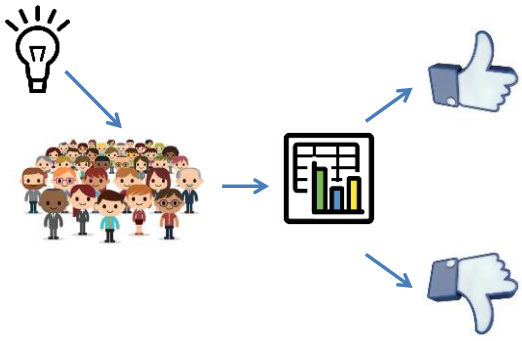
Experimental Pitfalls

Helen C. Purchase
University of Glasgow

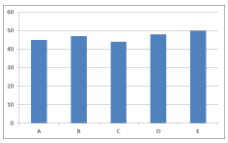
When we have an idea, we want to test it to see if it is a good one



Ideas that relate to the human use of computers need to involve participants in the testing



And after all the effort…



No experiment can ever be perfect

No experiment can ever be perfect

Conditions

Decisions

Experimental objects

Tasks

Nature of the participants

Number of participants

Allocation of conditions to participants

Location Equipment Online?

Experimental timing

Pre- and post-experimental activities

Data collection methods Data analysis methods

"Experimental Pitfalls"

or:

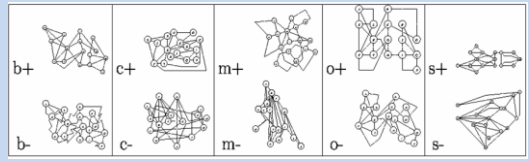
"Five Experimental Failures and a Joke"

or:

"Five Things I have Learned (and a Joke)"

Five Things I have Learned

… subject variability
… use of randomisation
… random factors
… piloting
… decision making



Robert Cohen and Murray James

"Validating Graph Drawing Aesthetics", Graph Drawing Symposium, 1996

Object-oriented class diagrams

Aesthetics:
bends, crosses, orthogonality, upward-flow

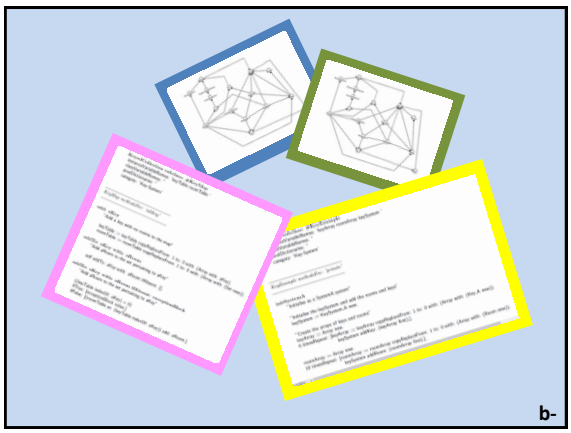
Eight conditions:
b+ b- c+ c- o+ o- f+ f-

Experimental object (program code):
System for storing information about keys and the doors they can unlock – two versions (distributed, centralised)

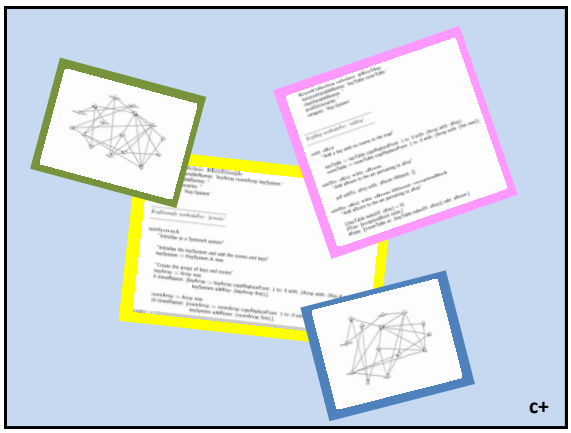
Steve Grunden

Unpublished, 1997

	Green (distributed)	Blue (centralised)
Few bends (b-)		
Many crossings (c+)		
Not much orthogonality (o-)		
Mostly upward direction (f+)		



b-



c+

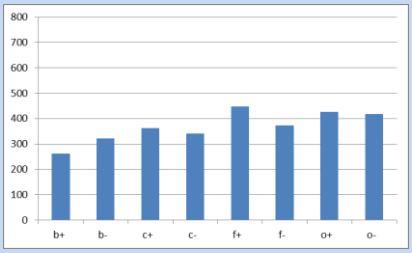
Green diagram
- □ Pink code
- □ Yellow code
- □ Neither

Blue diagram
- □ Pink code
- □ Yellow code
- □ Neither

The experiment

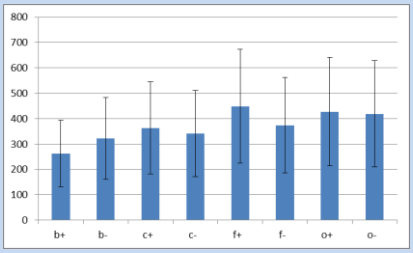
- Third-year computer science class
- Top quartile of the relevant class in mid-semester test
- N=49, one-to-one
- Between-participants (6 per condition + 1)

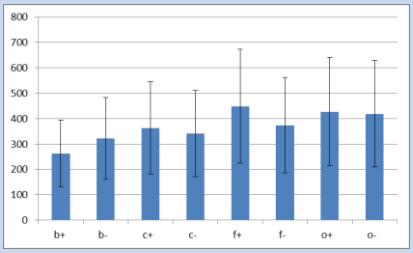
- Pre-experiment tutorial
- Data: time to get correct answer



Independent measures t-test *(b+/f+)*: p=0.078

N=49





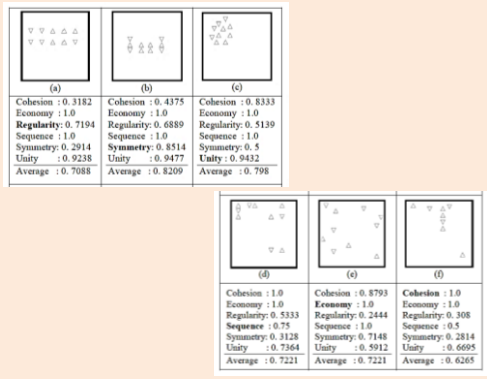
Subject variability!

So…

… it is hard to ensure equivalent domain knowledge in a between-participants' experiment

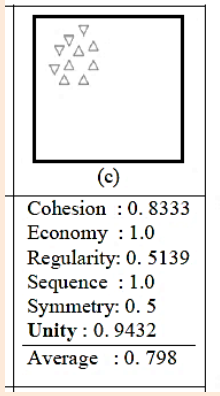
Screen Layout Principles	Good	Bad
$\text{Cohesion} \equiv \frac{\lvert CM_{tl}\rvert + \lvert CM_{lc}\rvert}{2} \in [1,0]$		
$\text{Economy} \equiv \frac{1}{n_{size}} \in [1,0]$		
$\text{Regularity} \equiv \frac{RM_{alignment} + RM_{spacing}}{2} \in [1,0]$		
$\text{Sequence} \equiv 1 - \frac{\sum_{j=UL,UR,LL,LR} \lvert q_j - v_j\rvert}{8} \in [1,0]$		
$\text{Sym} \equiv 1 - \frac{\lvert SYM_{vertical}\rvert + \lvert SYM_{horizontal}\rvert + \lvert SYM_{radial}\rvert}{3} \in [1,0]$		
$\text{Unity} \equiv \frac{\lvert UM_{form}\rvert + \lvert UM_{space}\rvert}{2} \in [1,0]$		

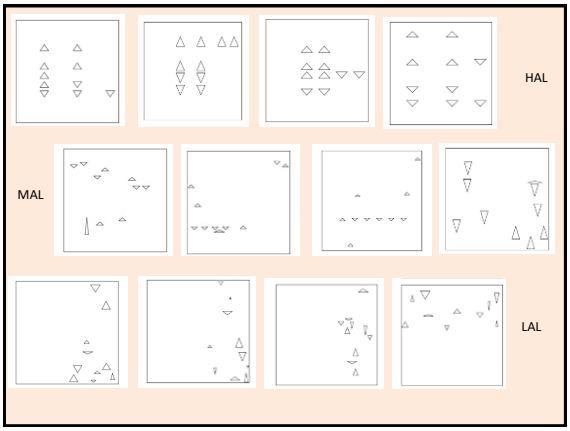
Ngo, et al (2003)

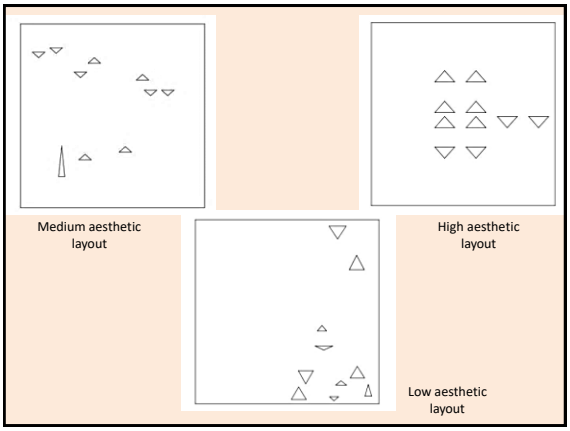


Carolyn Salimun

"The effect of aesthetically pleasing composition on visual search performance", Nordic HCI, 2010







90 stimuli

10 for each of nine aesthetic categories

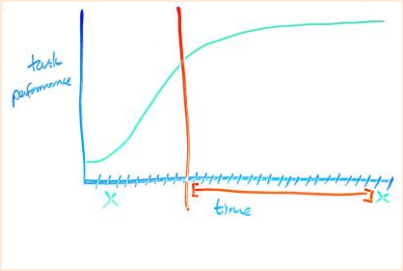
CATEGORY	LAYOUT METRICS					
	Cohesion	Economy	Regularity	Sequence	Symmetry	Unity
1. HAL	High	High	High	High	High	High
2. MAL	Medium	Medium	Medium	Medium	Medium	Medium
3. LAL	Low	Low	Low	Low	Low	Low
4. High cohesion	**High**	Low	Low	Low	Low	Low
5. High economy	Low	**High**	Low	Low	Low	Low
6. High regularity	Low	Low	**High**	Low	Low	Low
7. High sequence	Low	Low	Low	**High**	Low	Low
8. High symmetry	Low	Low	Low	Low	**High**	Low
9. High unity	Low	Low	Low	Low	Low	**High**

$0.7 \leq$ High ≤ 1.0, $0.5 \leq$ Medium < 0.7, $0.0 \leq$ Low < 0.5

9 conditions: 3 overall average aesthetic, cohesion, economy, regularity, sequence, symmetry, unity

N=21

- **RQ:** "Does layout aesthetic affect visual effort?"
- **Task:** count the number of upright triangles
- Within-participants experimental design
- **Dependent variables:** accuracy, response time, scan path length, scan path duration, number of fixations, fixation duration/gaze time
- **Independent variables:** aesthetics levels (high, medium, low), layout metrics



90 stimuli

- 10 practice tasks
- 90 stimuli => 90 factorial possible sequences

- 10 practice tasks
- 90 stimuli => 90 factorial possible sequences
- Only two used

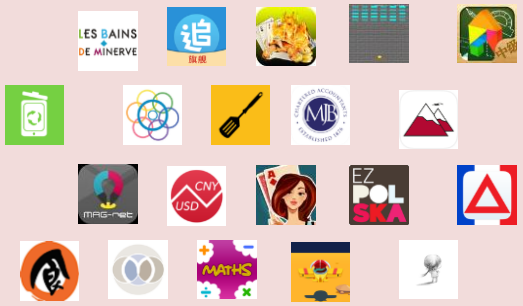
- 10 practice tasks
- 90 stimuli => 90 factorial possible sequences
- Only two used… discovered after all data collected
- Performance not analysed
- Focus analysis on eye movements

N=21

So…

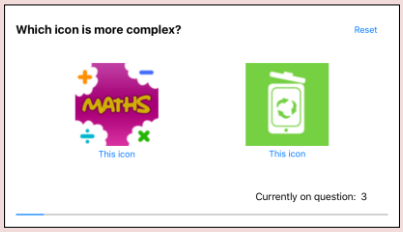
… randomisation introduces important variability that can mitigate against unwelcome learning effects

20 icons randomly chosen from the most recently launched apps on Google Play



Rory Bain

Unpublished student thesis, 2016



SubjectiveComplexity =

-0.024 + 0.000905col + 0.094hog

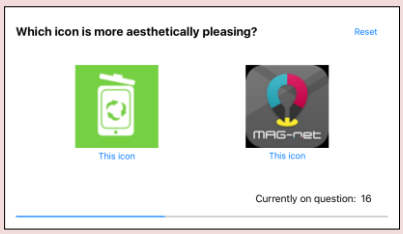
The "Histogram of Oriented Gradients" algorithm measures the number of distinct objects in an image

N=22

High Complexity

Medium Complexity

Low Complexity



SubjectiveAesthetics

= 0.659 – 0.0005926col – 0.001har

The "Harris Corner" algorithm measures the number of corners in an image

N=22

High Aesthetics

Medium Aesthetics

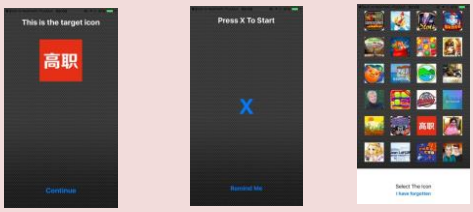
Low Aesthetics

RQ: does complexity/aesthetics affect search efficiency?
Task: icon search time

Complexity		Target icon		
		high comp	medium comp	low comp
Distractors	high complexity	*hh-c*	*mh-c*	*lh-c*
	medium complexity	*hm-c*	*mm-c*	*lm-c*
	low complexity	*hl-c*	*ml-c*	*ll-c*

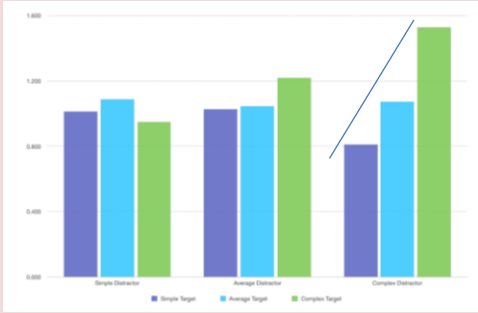
Aesthetics		Target icon		
		high aesth	medium aesth	low aesth
Distractors	high aesthetic	*hh-a*	*mh-a*	*lh-a*
	medium aesthetic	*hm-a*	*mm-a*	*lm-a*
	low aesthetic	*hl-a*	*ml-a*	*ll-a*

Does complexity affect search time?



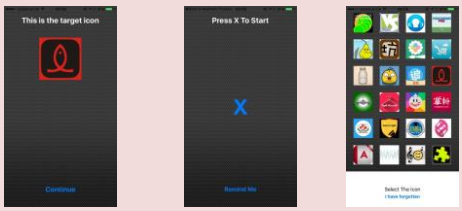
low complexity, with high complexity distractors

n=34



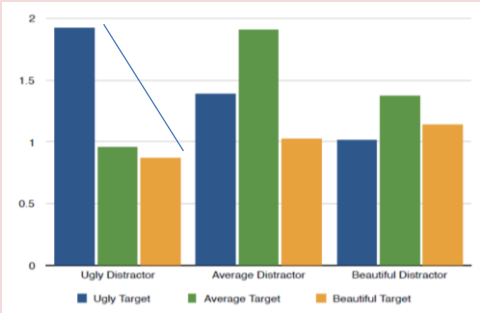
The complexity of the target icon only has a search time effect (simple is quicker) when the distractors are complex

Does aesthetics affect search time?



High aesthetic, with medium aesthetic distractors

n=34



The aesthetics of the target icon only has a search time effect (beautiful is quicker) when the distractors are ugly

The complexity of the target icon only has a search time effect (simple is quicker) when the distractors are complex

The aesthetics of the target icon only has a search time effect (beautiful is quicker) when the distractors are ugly

But what about other features of the icons we have not considered: e.g. elegance, or metaphoric association, or balance, or symmetry?

Complexity		Target icon		
		high comp	medium comp	low comp
Distractors	high complexity	*hh-c*	*mh-c*	*lh-c*
	medium complexity	*hm-c*	*mm-c*	*lm-c*
	low complexity	*hl-c*	*ml-c*	*ll-c*

Third dimension: symmetry

symmetry

Complexity		Target icon		
		high comp	medium comp	low comp
Distractors	high complexity	*hh-c*	*mh-c*	*lh-c*
	medium complexity	*hm-c*	*mm-c*	*lm-c*
	low complexity	*hl-c*	*ml-c*	*ll-c*

symmetry

Complexity		Target icon		
		high comp	medium comp	low comp
Distractors	high complexity	*hh-c-ns*	*mh-c-ns*	*lh-c-ns*
	medium complexity	*hm-c-ns*	*mm-c-ns*	*lm-c-ns*
	low complexity	*hl-c-ns*	*ml-c-ns*	*ll-c-ns*

High symmetry
Medium symmetry
Low symmetry
No symmetry

Fourth dimension: metaphoric association…

Control/ Random/ Confounding factors

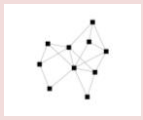
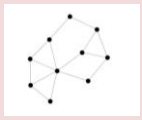
Conditions: deliberately change

Control: deliberately ensure that they don't change

Random: deliberately allow to change randomly to ensure generalisability

Confounding: factors that change together with the conditions (even though you don't want them to)

I. Scott McKenzie, Human-Computer Interaction, 2013


Conditions: shapes of nodes *(square, triangle, circle)*

Control: deliberately ensure that they don't change *(number of nodes)*

Random: deliberately allow to change to ensure generalisability *(density)*

Confounding: factors that change together with the conditions *(graph drawings with triangular nodes have longer edges)*

Conditions: effectiveness of two biological diagrams for learning (A and B)

Control: deliberately ensure that they don't change *(first year biology students)*

Random: deliberately allow to change to ensure generalisability *(age)*

Confounding: factors that change together with the conditions *(the biology students who study chemistry performed better than those who didn't)*

So…

… we need to think carefully about what we *want* to control, *can* control, *can't* control, and *don't care about* controlling

Scrolling Behaviour with Single- and Multi-column Layout

Comparing:

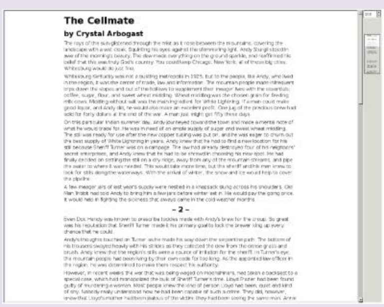
- Vertical scrolling: single column (web browsers)
- Horizontal scrolling text: multiple columns of same height (electronic readers)

How do people read?

How do people scroll?

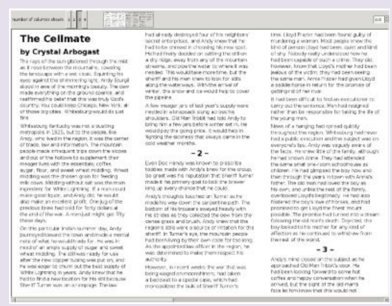
Anonymous

Unpublished 2009



The Cellmate

by Crystal Arbogast

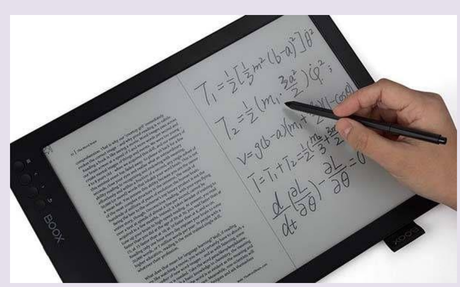


The Cellmate

by Crystal Arbogast

"Our results suggest that horizontal-scroll layout will be particularly popular on devices such as e-book readers that have slow display refresh and so are not well-suited to continuous scrolling…

….We plan to conduct further studies to see if our findings generalize to other kinds of participants, devices and reading material."



Experimental process

- Demo of device
- Condition A
  - demo and training
  - read story, answer three simple questions
- Condition B
  - demo and training
  - read story, answer three simple questions
- Data:
  - Logging (eye-tracking)
  - Preference questionnaire

From my notes (verbatim)

- P1: One stylus is not enough
- P2: Problem with vertical scrolling using the stylus directly on the text – text jumps DOWN a little before moving UP
- P3: Definitely is a problem with the vertical scrolling
- P5: System hung during horizontal training (totally unresponsive). Reset. System crashed during reading of HR. Reset and put charger in. System crashed again near the end of reading HR. Reset. Crash during the HR questions. Experiment abandoned
- P8: rapid, uncontrollable scrolling
- P9: a problem with sticking buttons

38 students had been recruited in advance:
"I'm afraid that I am going to have to cancel our experimental session next week - the mobile device we use has unexpectedly developed a fault."

So…

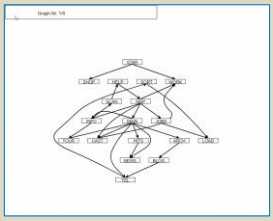
… never (ever, ever) remove the piloting step!

Dynamic graphs

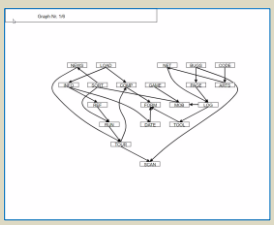
- "Does maintaining the 'mental map' help in understanding evolving graphs?"
- Conditions:
  - low mental map
  - medium mental map
  - high mental map
- Three different evolving graphs
  - 14-20 nodes, 15-30 edges, 4 changes/time-slice
- Four different tasks
  - addition/removal of edges, overall structure

Eve Hoggan and Carsten Görg

How Important Is the "Mental Map"?
Graph Drawing Symposium, 2005



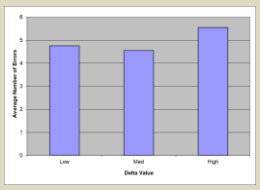
Graph 1: low mental map
(lots of movement)



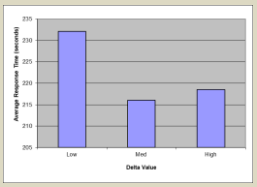
Graph 3: high mental map
(minimal movement)

n=20

Aggregating over all three graphs and all four questions

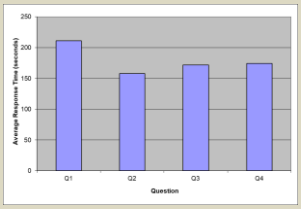


Errors: no significant difference

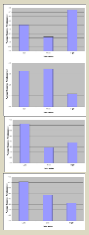


Response time: no significant difference

Individual questions



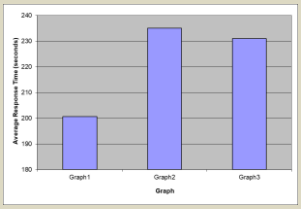
As expected, the questions were of different difficulty



Q1
Q2
Q3
Q4

Q1: number of new edges
Q2: node with most changes
Q3: year of extreme reduction in size
Q4: year a particular page had degree one

Individual graphs



We did not expect the graphs to be of different difficulty

"We aimed to keep the size and changes of these graphs as similar as possible, while keeping them distinctive…having made an effort to keep the three evolving graphs comparable (similar size, similar number of changes per time-slice)…"

Characterisation of tasks & objects

- It was easy to identify the difference between the tasks in terms of difficulty – and so justifiable to analyse the data according to task
- It was impossible to identify any difference between the graphs…because they had been arbitrarily defined

So…

… never make arbitrary decisions (they may come back to haunt you!)

Five Things I have Learned

- … it is hard to ensure equivalent domain knowledge in a between-participants' experiment
- …randomisation introduces important variability that can mitigate against unwelcome learning effects
- …we need to think carefully about what we want to control, can control, can't control, and don't care about controlling
- …never (ever, ever) remove the piloting step!
- …never make arbitrary decisions (they may come back to haunt you!)

The identification of groups in social networks drawn as graphs is an important task for social scientists who wish to know how the population divides with respect to relationships or attributes …. In this paper, we report on an experiment … We find that, despite the use of colour as the pre-attentive visual feature to signify group membership, participants tend to rely on structure as the basis for their visual community identification.

6th February 2020

The identification of groups in social networks drawn as graphs is an important task for social scientists who wish to know how the population divides with respect to relationships or attributes …. In this paper, we report on an experiment … We find that those algorithms that clearly separate communities with large distances are most effective, while the use of colour to represent community membership is more successful than reliance on structural layout.

13th February 2020

In summary…

- Experiments are fun…
- …but time-consuming, difficult, and can never be perfect
- *Every* decision counts
- We are all still learning…
- …and it is often easier to imitate others' processes than consider whether they are really appropriate