Computational Behavioural Science

Bayesian Cognitive Models

Bayesian Models

Bayesian models are widely used in cognitive science, especially in studying human methods of causal inference and categorisation. These are typically not process models – they don't try to capture the cognitive steps people go through when they solve problems. Instead, they make predictions about what sort of judgements people tend to make in particular types of problems.

Key Bayes formulae

- Marginalisation: $P(a) = \sum_b P(a, b)$
- Conditional probability: $P(a|b) = \frac{P(a,b)}{P(b)}$
- Chain rule: $P(a|b)P(b) = P(a, b)$
- Baes rule: $P(a|b) = \frac{P(b|a\,)P(a)}{P(a)}$

Bayesian Concept Formation

Let $X = \{x_1, x_2, \ldots x_n\}$ be a set of examples of a concept C. Here we will use numbers drawn from the interval $[0,100]$. This is based on a model of concepts developed by Josh Tenenbaum.

Let H by the Hypothesis space of possible concepts. Here it is all possible subsets of the interval 0 to 100. In his model, Tenenbaum considers the following hypotheses:

1. Mathematical properties (24 hypotheses):
    a. Odd, even, square, cube, prime numbers
    b. Multiples of small integers
    c. Powers of small integers
2. Raw magnitude (5050 hypotheses):
    a. All intervals of integers with endpoints between 1 and 100.

The total probability assigned to mathematical concepts is λ, while the total probability assigned to magnitude concepts is 1 − λ. Within each category of concepts, a uniform prior is used.

In strong sampling, it is assumed that the data are intentionally generated as positive examples of a concept, while in weak sampling, it is assumed that the data are generated without any restrictions.

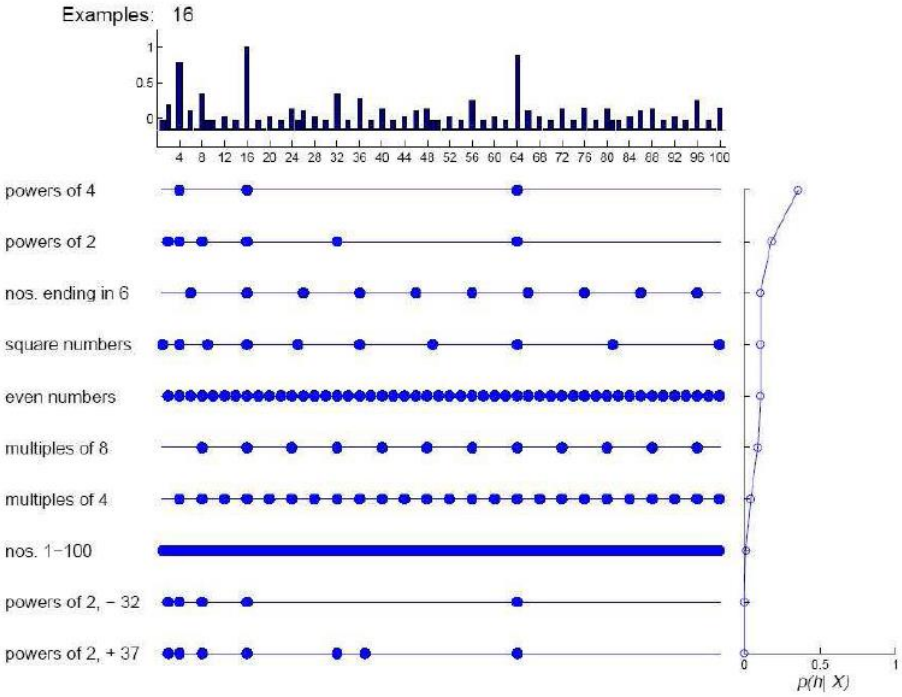
In **strong sampling**, we assume observation is randomly sampled from the true hypothesis:

$$P(x|h) = \begin{cases} \frac{1}{|h|} & , \text{if } x \in h \\ 0 & , \text{otherwise} \end{cases}$$

In **weak sampling**, we assume observations randomly sampled and then classified:

$$P(x|h) = \begin{cases} 1 & , \text{if } x \in h \\ 0 & , \text{otherwise} \end{cases}$$

Some examples of this likelihood are shown below:



When considering a new item y, we must consider the set of all remaining hypotheses consistent with X and y, which we can denote H_y. We then have:

$$P(y \in C|X) = \sum_h P(y \in C, h|X)$$

$$= \sum_h P(y \in C|h, X)P(h|X)$$

$$P(y \in C|X) = \sum_h P(y \in C|h)P(h|X)$$

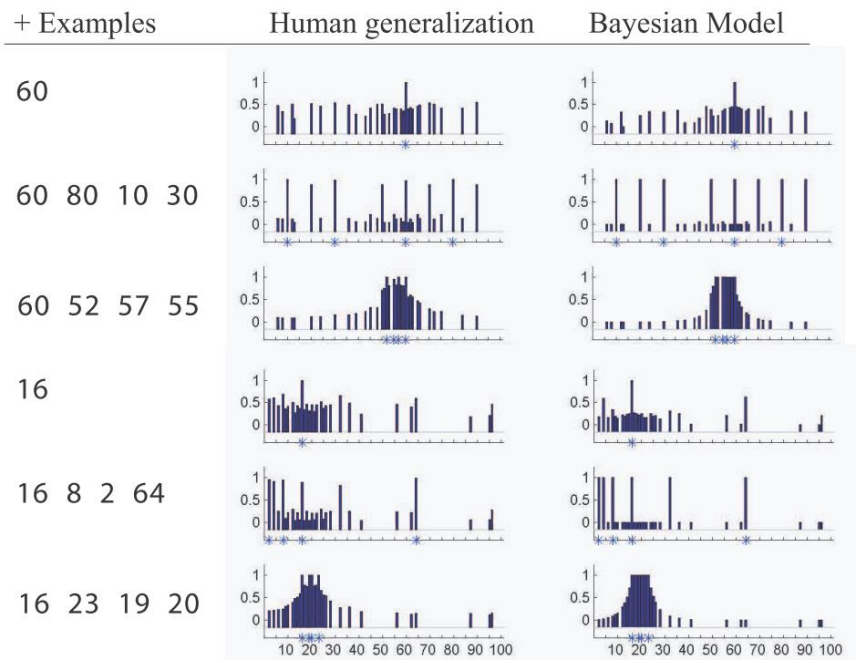
Since $P(y \in C|h) = 1$ if $h \in H_y$ and zero otherwise, we can simplify the above equation to:

$$P(y \in C|X) = \sum_{h \in H_y} P(h|X)$$

$$P(y \in C|X) = \sum_{h \in H_y} \frac{P(X|h)P(h)}{\sum_{h' \in H_y} P(X|h)P(h')}$$

This method is known as hypothesis averaging.

As shown in the figure below, this method produces very similar generalisations to human subjects.

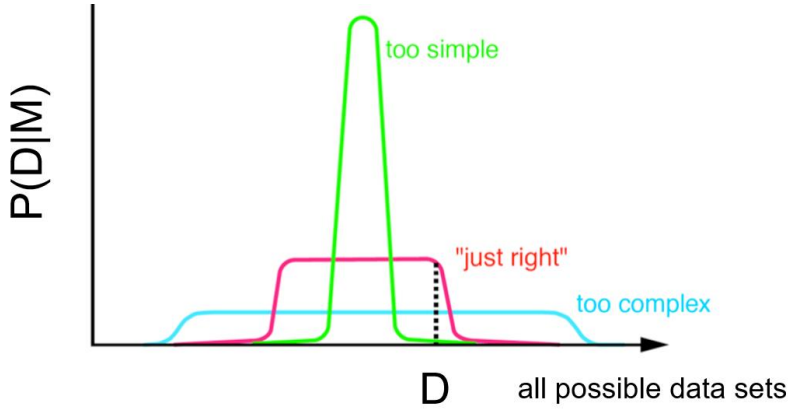


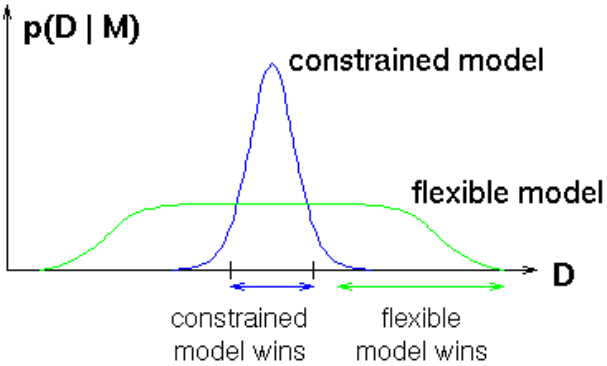
Bayesian Model Selection

Using observed data to choose between two probabilistic models that differ in their complexity is often called the problem of model selection. Complex hypotheses have more degrees of freedom that can be adapted to the data, and can thus always be made to fit the data better. The Bayesian method for model selection holds that the model should be chosen that has the maximum posterior probability given the data, averaging over all possible parameters of the model. This is computed as:

$$p(m|d) = p(d|m)p(m)$$

$$p(m|d) = \int p(d|m,\theta)p(\theta|m)\,d\theta \times p(m)$$





Bayesian Networks

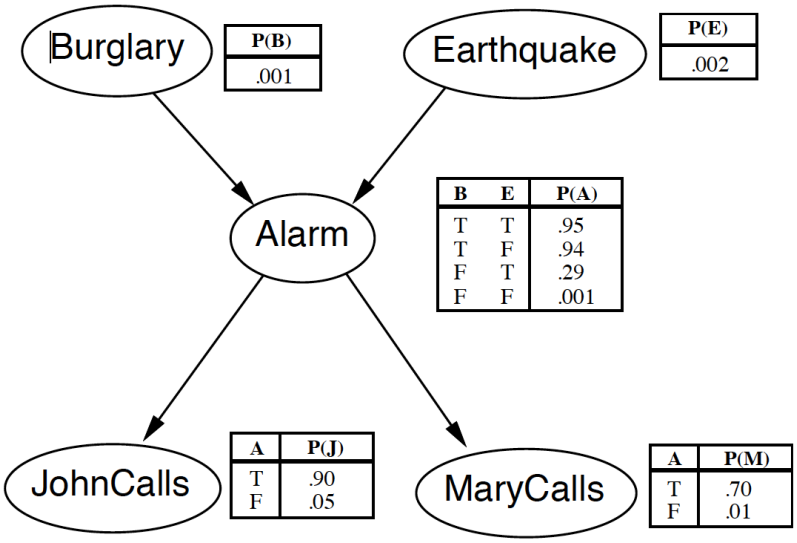
Standard Bayes Networks

A Bayesian network specifies a joint probability distribution such that each variable corresponds to a node in the network, and an edge connection from node i to node j means that X_j is independent of all other nodes, conditional on X_i (and any other parent nodes). Also, each node as a conditional probability distribution that specifies how that node depends on the values of only its parent nodes $Pa(X_i)$. Thus we can represent the joint distribution as:

$$P(X_1, X_2, \dots, X_n) = \prod_I P\left(X_i | Pa\ (x_i)\right)$$

Bayesian networks have several advantages:

- Bayesian networks help modelers define high dimensional distributions.
- Bayesian networks provide a concise way of representing probability distributions.
- Bayesian networks are modular and therefore easy to extend.
- Bayesian networks often support efficient inference.



Note that different Bayes nets can capture the same joint distribution, because a given joint distribution can be factorised in different ways. For example:

$$P(a, b) = P(a|b)P(b)$$
$$P(a, b) = P(b|a)P(a)$$

Markov Chain Methods

Markov Chain Monte Carlo (MCMC) simulation is a way to draw samples from a distribution. The approach relies on a Markov Chain that specifies transitions between values of x. The Gibbs sampling and the Metropolis-Hastings algorithm are two examples of MCMC methods. Gibbs sampling allows us to sample from a probability distribution over multiple variables if all we can easily do is sample from the conditional distribution of each component on the others. This works as follows:

1. Choose $X^i = \left(x_1^i, x_2^i, x_3^i\right)$.
2. Take draws from the conditional distributions:
    a. $P\left(x_1^{i+1} | x_2^i, x_3^i\right)$

b. $P\left(x_2^{i+1} | x_1^{i+1}, x_3^i\right)$
c. $P\left(x_3^{i+1} | x_1^{i+1}, x_2^{i+1}\right)$

3. Repeat k times.

Ignore the first n set of draws as the burn-in and keep the remaining $k - n$ draws. The result is a set of draws from the joint distribution of X.

Causal Bayes Networks

The principal of common cause:

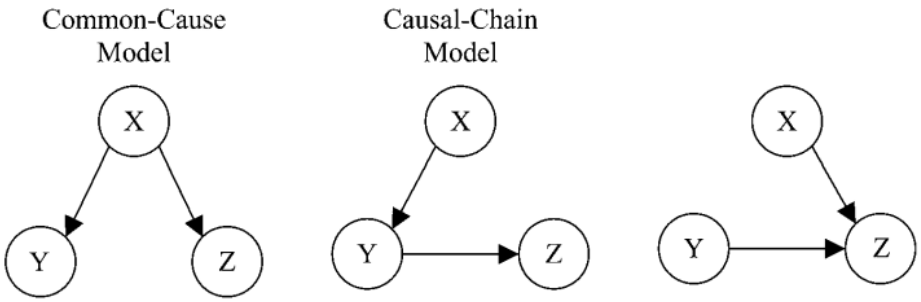
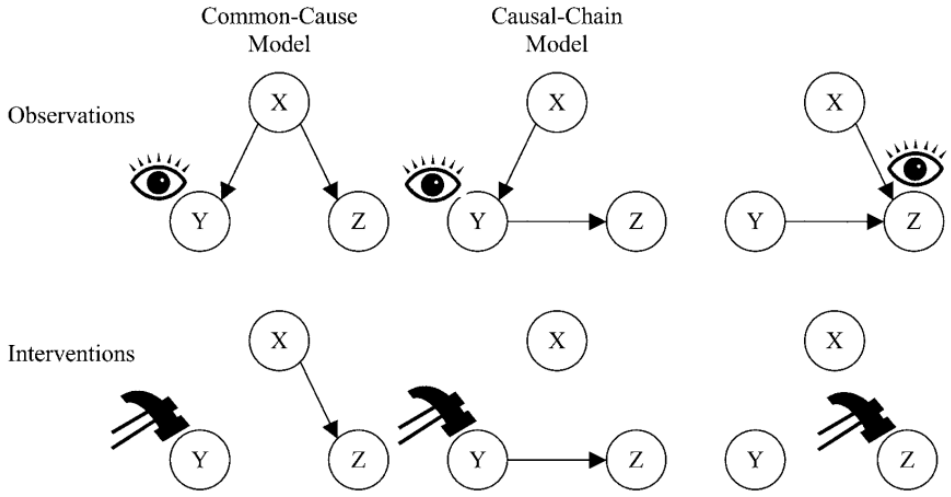


FIGURE 6-1 Three basic causal models.

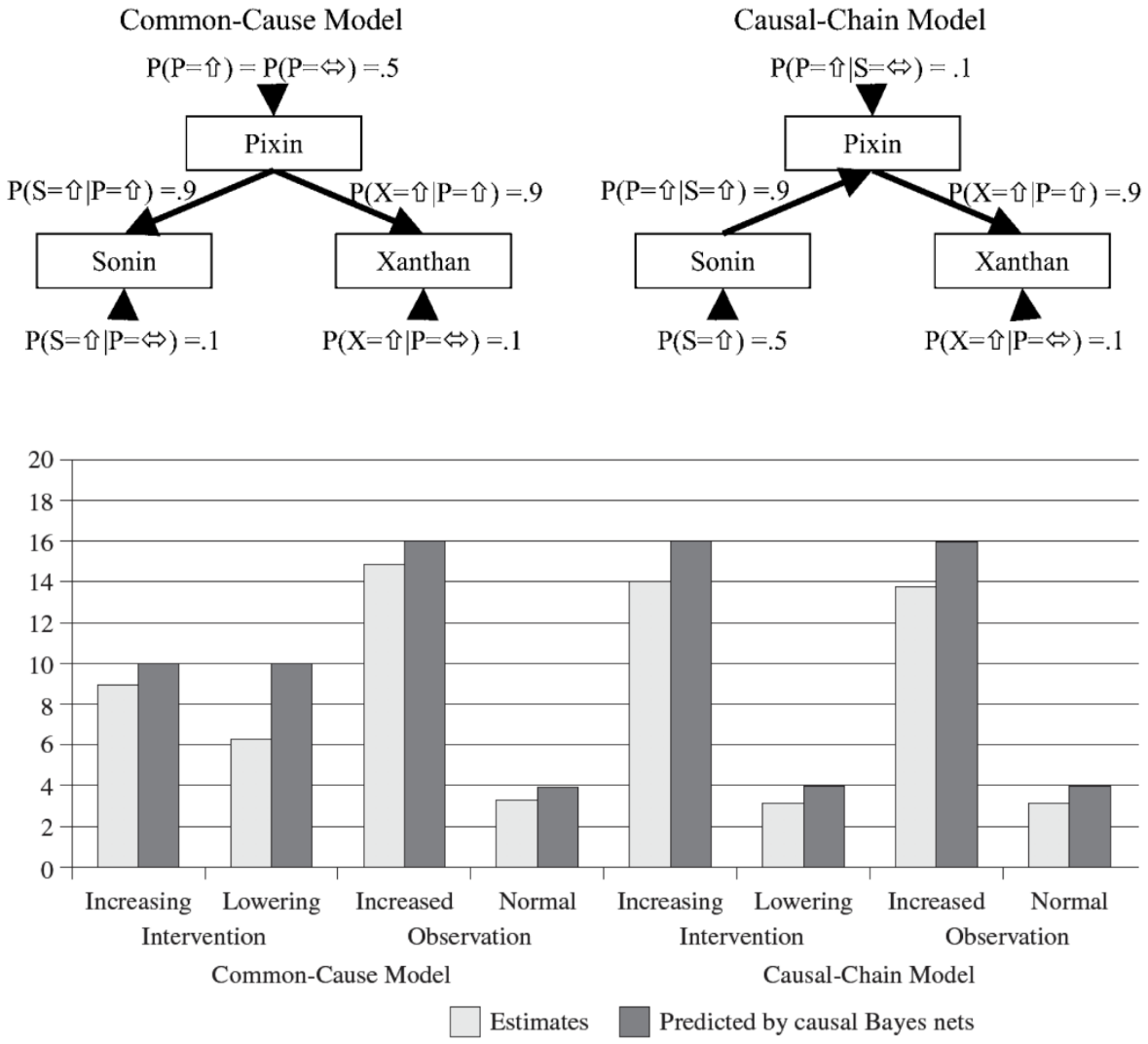
Traditional Bayes nets and other probabilistic theories are incapable of distinguishing between observations and interventions because they lack the expressive power to distinguish observational and interventional conditional probabilities. Both types are subsumed under the general concept of conditional probability. Causal Bayesian networks explicitly represent causal relationships, not just probabilistic dependencies. Since the edges have a causal interpretation, we can reason about interventions. This is possible because of the introduction of the 'do' operator. The 'do' operator (symbolised below as a hammer) sets the value of a node to some specified value, and cuts any link that node has with its parents (because such dependencies are destroyed by the intervention). Whereas observations leave the surrounding causal network intact, interventions alter the structure of the causal model by rendering the manipulated variable independent of its causes.



For example, the bottom left intervention can be represented as follows:

$$P(X, do(Y = 1), Z) = P(Z|X)P(X)$$

Causal Bayes Networks can predict the way that humans distinguish between merely correlation relations and causal relations. An example of such experimental results is shown below.

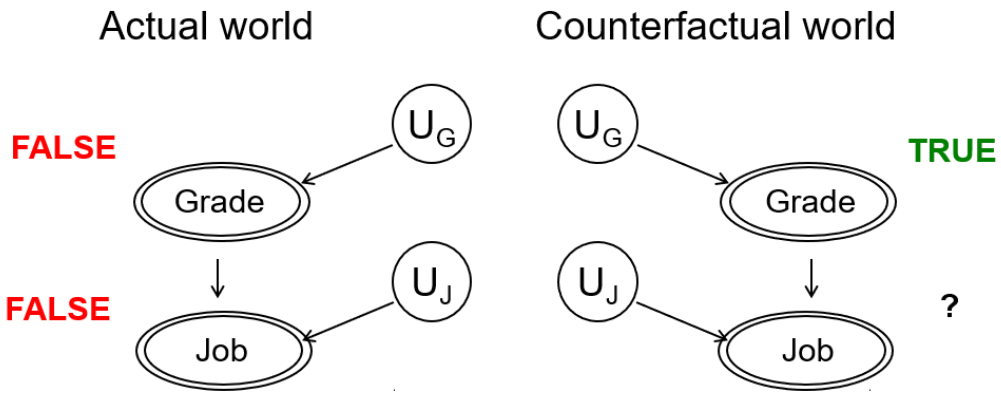


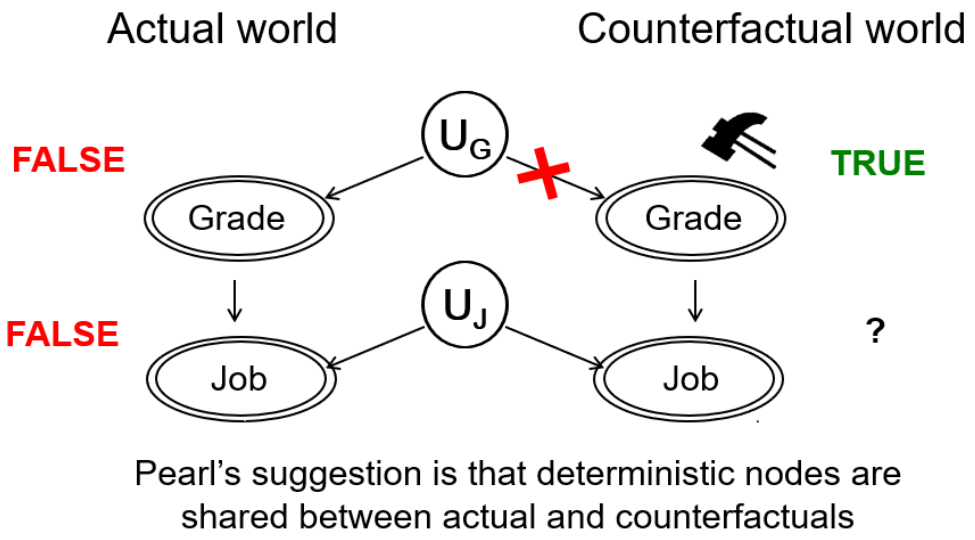
Counterfactual Bayes Networks

A counterfactual is a unique form of inference in which A->B as interpreted to mean that B would be true if A were made true by an intervention. Counterfactuals pose unique problems for models of Bayesian inference, since they are typically understood to be 'non-backtracking, meaning that one does not reason backwards from a counterfactual supposition to draw conclusions about the causes of the hypothetical situation. For instance, one would not reason "if the meat had been cooked rare, then the flame would have been set to low". Modus ponens fails in such models, since just because A is true in the actual world, and if A were set by intervention to be true then B would follow, it doesn't follow that B is true in the actual world. We can accommodate this form of reasoning using a type of graph called a functional causal model (FCM).

In FCMs, all nodes are deterministic except for the exogenous nodes (shown below as single circles). Compared to generic causal networks, FCMs are more precise about the underlying causal mechanisms. FCMs use a 'twin network' structure, with one half corresponding to the actual work, and the other half to the counterfactual world. In the figure below, the exogenous nodes U_G and U_J respectively represent whatever underlying mechanisms are responsible for an applicant either

having good grades or not, and getting a job or not. It is important to understand that once these exogenous circular nodes are set, everything else in the network is deterministic. This fits the intuition that in the counterfactual world, we should not get 'another roll of the dice' for any stochastic variable when moving to the counterfactual world. In other words, instantiating a counterfactual event is equivalent to an imaginary intervention on a causal model in which all variables that are not affected by the intervention are assumed to stay at currently observed levels. An intervention in such a network corresponds to cutting the link between one of the exogenous nodes (below U_G) and its children in the counterfactual world, and then equating the values of the exogenous variables in both worlds.





The following table summarises what types of inferences are permissible under different types of Bayes nets.

	Bayes net	Causal Bayes net	Functional causal model
Observations	✓	✓	✓
Interventions		✓	✓
Counterfactuals			✓

Causation and Categorisation

Causal Structure Learning

	Effect present (e^+)	Effect absent (e^-)
Cause present (c^+)	$N(e^+, c^+)$	$N(e^-, c^+)$
Cause absent (c^-)	$N(e^+, c^-)$	$N(e^-, c^-)$

The ΔP Model is a successful and widely used model of causal relationships. In this model the inferred strength of causal relationship is:

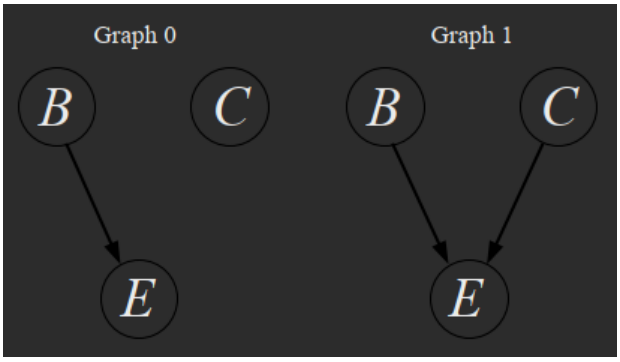
$$\Delta P = \frac{N(e^+, c^+)}{N(e^+, c^+) + N(e^-, c^+)} - \frac{N(e^+, c^-)}{N(e^+, c^-) + N(e^-, c^-)}$$

An alternative model called causal power is a modification of ΔP:

$$power = \frac{\Delta P}{1 - P(e^+|c^-)}$$

Yet another method known as 'causal support' uses the relative probability of the data under two alternative Bayesian causal networks.

$$support = \log \frac{P(d|Graph\ 1)}{P(d|Graph\ 0)}$$

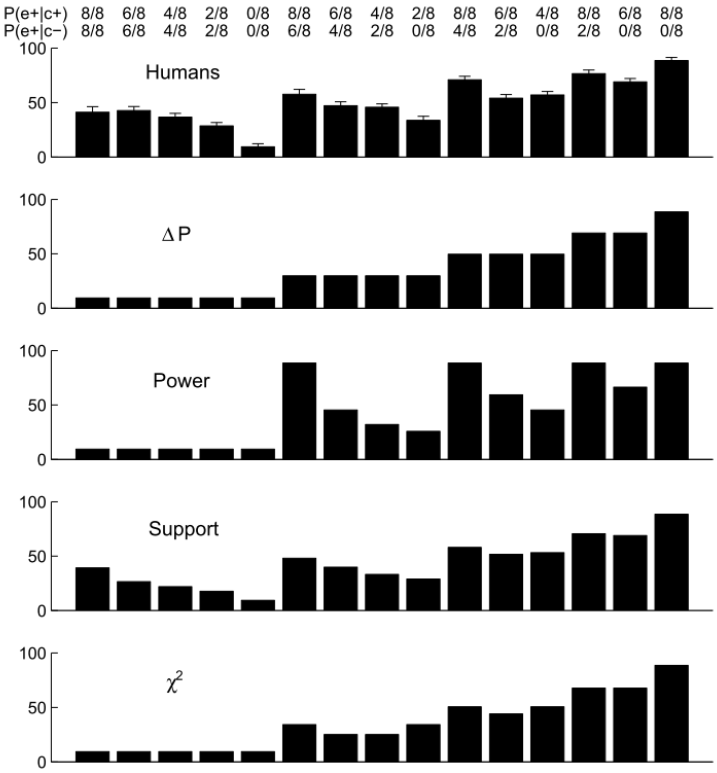


where $P(d|\text{Graph 1})$ and $P(d|\text{Graph 0})$ are computed by integrating over the parameters associated with the different structures

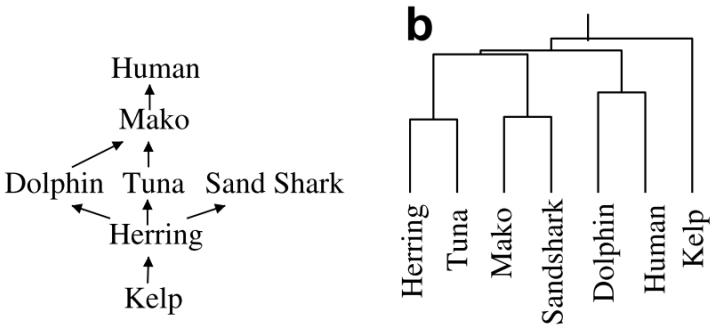
$$P(d|\text{Graph 1}) = \int_0^1 \int_0^1 P_1(d|w_0, w_1, \text{Graph 1})\ P(w_0, w_1|\text{Graph 1})\ dw_0\ dw_1 \quad (24)$$

$$P(d|\text{Graph 0}) = \int_0^1 P_0(d|w_0, \text{Graph 0})\ P(w_0|\text{Graph 0})\ dw_0. \quad (25)$$

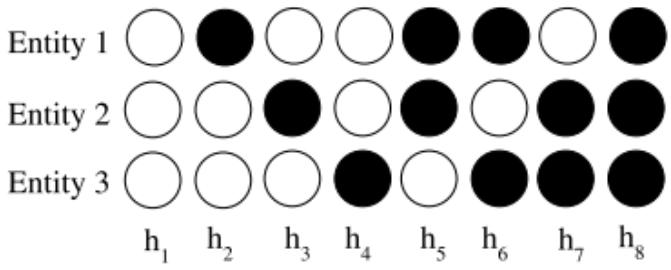
The causal support method provides a very close fit to human judgements in at least one type of task, as shown in the figure below.



When making judgements about the spread of a disease through a food chain, humans tend to incorporate taxonomic information (shown in right), as well as the distance between nodes (causal distance), and an asymmetry between prey versus predator (causal asymmetry).



These results can be very accurately predicted by a Bayesian network with a set disease transmission probability. We then enumerate all possible extensions of each predicate (in this case ‘has disease’), as shown in the diagram below.



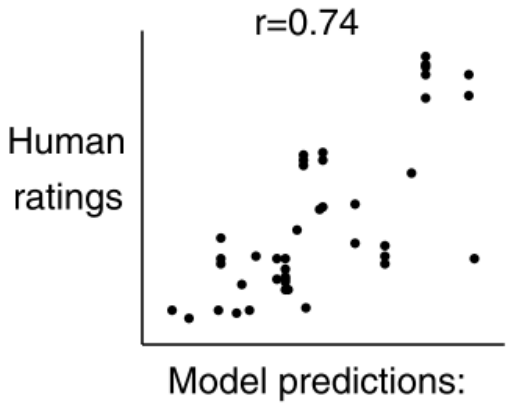
To make inference, we calculate the generalization probability $p(y|D)$ as equal to the proportion of hypotheses consistent with the data D that also include y, where each hypothesis is weighted by its prior probability $p(h)$:

$$p(y|D) = \frac{\sum_{y,D\in h} p(h)}{\sum_{D\in h} p(h)}$$

The prior for each hypothesis $p(h)$ is found by randomly assigning a subset of animals the disease, and seeing which other animals in the chain then acquire the disease. The prior probability of any hypothesis is equal to the proportion of times it ends up as the correct one (that is, when that hypothesis describes the distribution of the predicate in the chain).

$$p(h) = \frac{N_h}{N}$$

This model well predicts the judgments made by humans in this task.

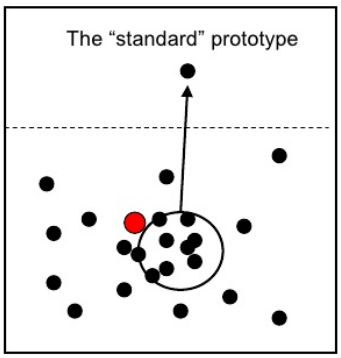


Supervised Categorisation

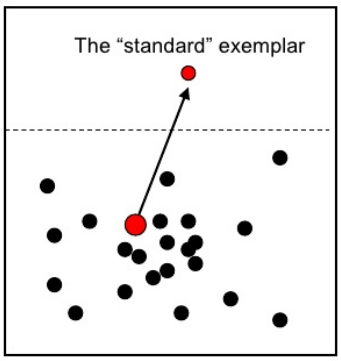
A concept is a mental representation of a class or individual that deals with what is being represented and how that information is typically used. It is common to distinguish between a concept and a category. A concept refers to a mentally possessed idea or notion, whereas a cate gory refers to a set of entities that are grouped together. A category is something in the world (e.g. the set of things that qualify as birds). A concept is something in the head (e.g. the mental representation of our knowledge of birds).

Three proposals about concepts

1. Rule-based approach: Concepts are rules (e.g. bachelor = unmarried and male).
2. Prototype models: A concept is a prototype that captures characteristic features.
3. Exemplar models: A concept is merely a set of stored examples.



A prototype is an abstract representation that is derived from the “center of mass” of the features of all the objects in the category.



An exemplar is a concrete representation. A specific instance that happens to be the most active in memory. More typical members tend to be the most active, on average.

A prototypical task is supervised categorisation. In supervised problems the learner observes a data set D of labeled examples (x_i, l_i) and must infer the category label l for a new object x_{new}. If we consider the simple case of two possible categories, the problem can be represented as:

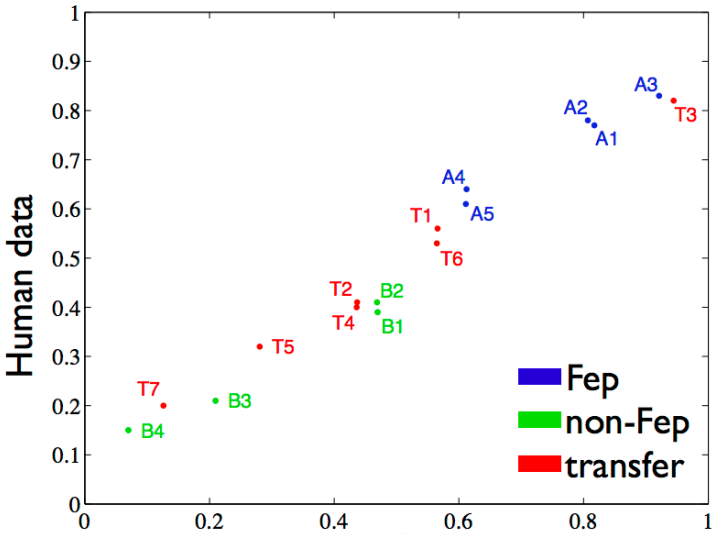
$$p(l_{new} = 1|x_{new}, D) = p(x_{new}|l_{new} = 1, D)p(l_{new} = 1|D)$$

The three models of categorisation then provide different rules for determining the likelihood:

1. Rule-based model: the likelihood is characterized by a rule for category k, and learn this rule via Bayesian inference.
2. Prototype model: the likelihood is a Gaussian distribution, and learn the mean and covariance for the distribution.
3. Exemplar model: use kernel-density estimation to compute the likelihood.

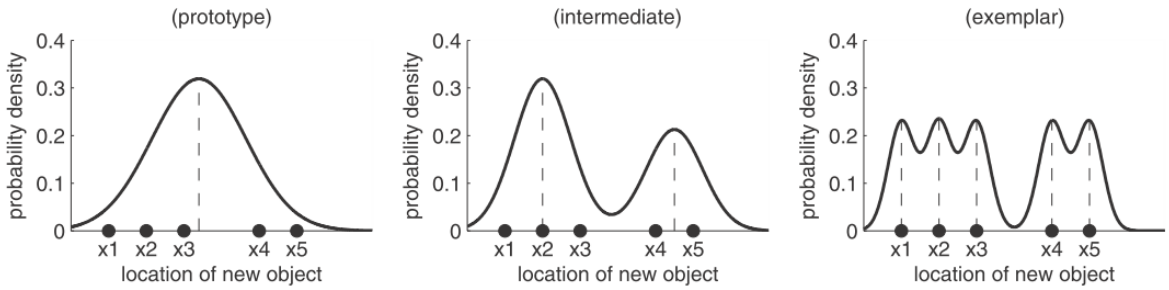
In some cases, rule-based methods (shown in figure below) are able to predict human performance very accurately. This is done by calculating the likelihood component of the above equation through summing over all rules r_k that are consistent with the data D:

$$p(x_{new}|l_{new} = 1, D) = \sum_{r_k} p(x_{new}|l_{new} = k, r_k)p(r_k|D)$$



However, rule-based approaches are limited in their applicability, since many concepts are graded and do not admit of any necessary and sufficient sets of rules. Humans also show typicality effects, whereby they consistently nominate certain items as better examples of their category than others, and take longer to make classification decisions in borderline cases.

Given these limitations of rule-based methods, much recent work has focused on the competing prototype and exemplar models. These are actually variations on a theme, with prototype models corresponding to parametric models clustered around a centroid, and the exemplar models corresponding to non-parametric kernel density estimation techniques.



According to the generalized context model (GCM) of classification, people represent categories by storing individual exemplars in memory, and classify objects based on their similarity to these stored exemplars. In GCM the similarity measure is context-sensitive, so the weights can 'stretch' the psychological space along highly attended, relevant dimensions, and to 'shrink' the space along unattended irrelevant dimensions.

A more sophisticated view about categories holds that it is not the similarity between an instance and the category that determines the instance's classification. Rather, it is the fact that our category provides a theory that explains phenomena in the world.

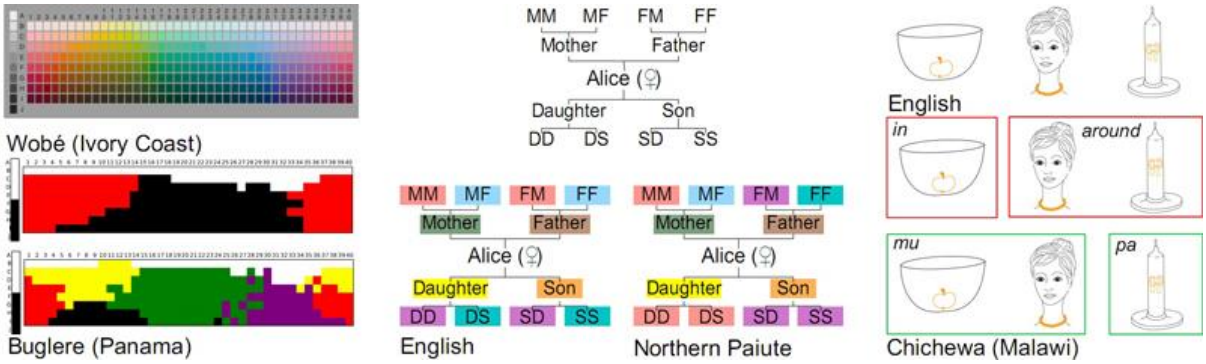
It should be noted that all of these models assume an underlying set of features or dimensions. The models don't explain where these come from, but methods for learning them include neural networks, brain imaging, or human labelling of automated clusters of images.

It seems that all of these models are useful in different contexts. Evidence for rule-based categories tends to be found with categories that are created from simple rules. Evidence for prototypes tends to be found for categories made up of members that are distortions around single prototypes. Evidence for exemplar models is strong when categories include exceptional instances that must be individually memorized. Evidence for theories is found when categories are created that subjects already know something about.

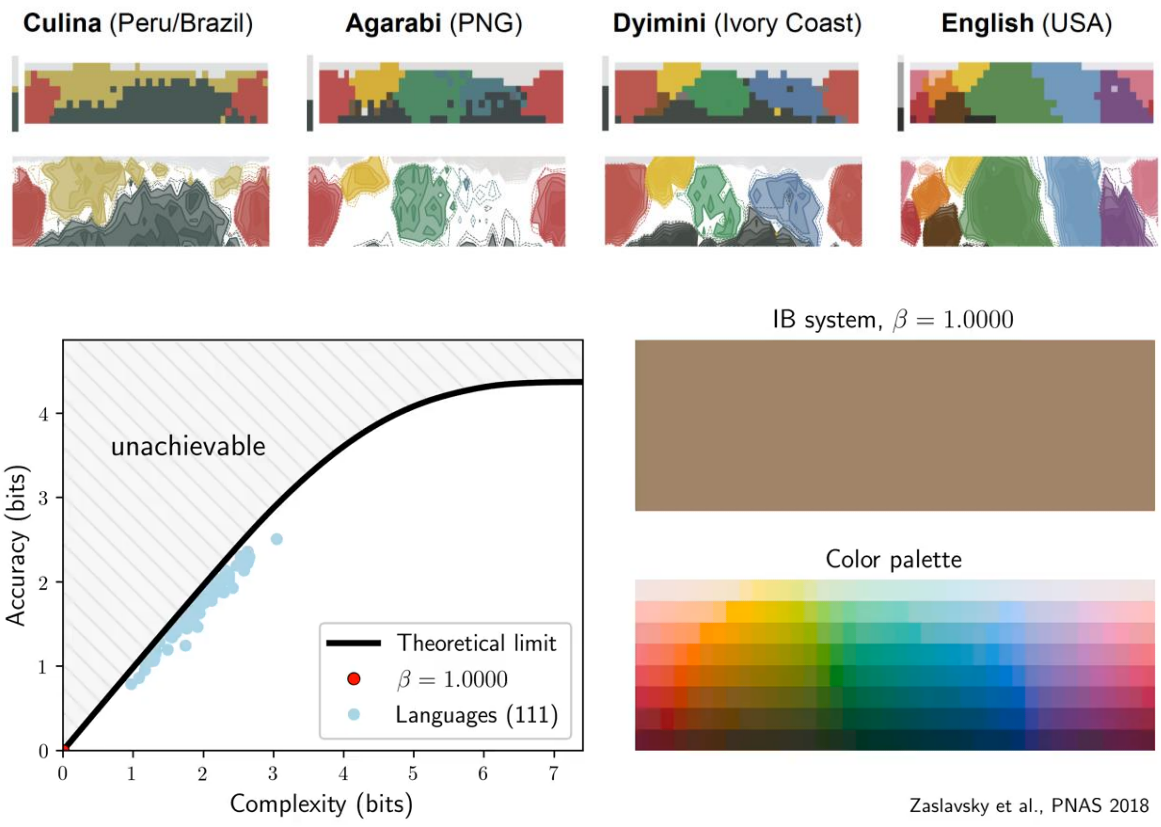
Unsupervised Categorization

In an unsupervised problem the learner observes unlabelled examples and must group them into categories (also called clusters).

Such classifications differ culturally in many domains.



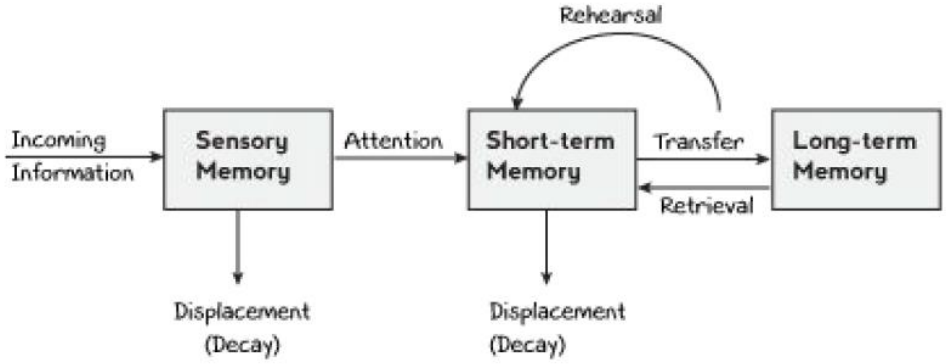
One study looking at colour words in different languages found that, although different languages had different numbers of colour terms, most languages lay very close to the boundary of near-optimal trade-offs between informativeness and complexity.



Episodic Memory

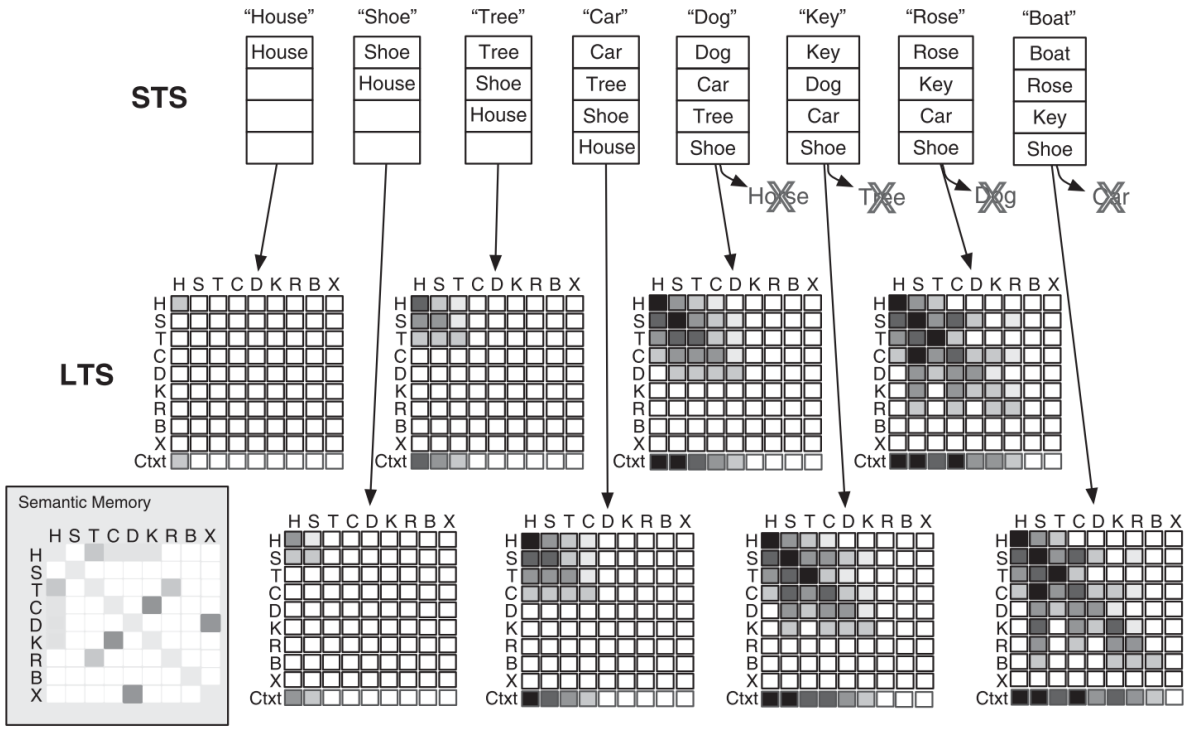
Modelling Episodic Memory

Multi-store models of memory distinguish between sensory, short-term, and long-term memory, and attempt to determine the mechanisms by which information is first encoded in and later recalled from, one form of memory to another.



The modern form of such multi-store models is known as the Search of Associative Memory (SAM), which models the exchange of information between the short-term store (STS) and long-term store (LTS). LTS is represented as a matrix containing values for the strengths of the associations formed through rehearsal, including pairwise associations among the list words, as well as associations between each list word and the list context. List context is conceptualized as the temporal and situational setting for a particular list.

Items enter STS in the order they are presented, while the participant rehearses the items occupying STS at any given time, thereby increasing the strengths of the items' episodic associations in LTS. When STS becomes full, each new item displaces one of the old items then occupying STS. Note that STS items are NOT displaced in the order they arrived, but rather according to a formula in which the probability of being displaced increases with time spent in the STS.



For each unit time an item spends in the STS, it increases its association with:

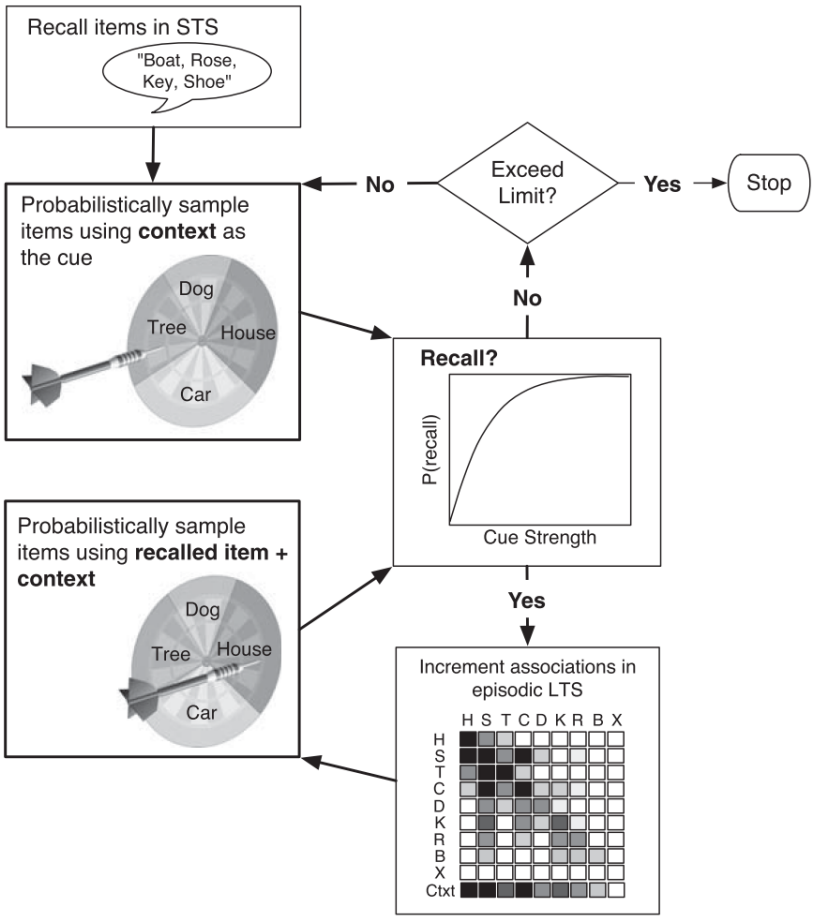
- the list context
- the next item in the STS
- the previous item in the STS (but by a smaller amount)
- its own unit (known as self-strength)

Retrieval of items from the LTS is a two-stage process in SAM. First the item must be 'sampled', which can be thought of as a subconscious rise in activation of that item, and then it must be 'recalled', which is the process of becoming consciously aware of that item. The probabilities for sampling and recalling an item are:

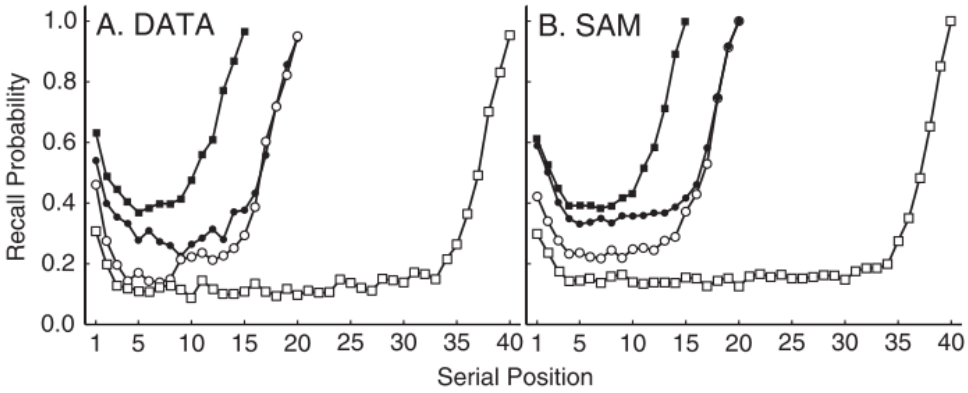
$$P_{\text{SAMPLING}}(i|j, \text{context}) = \frac{S(i,j)^{W_e} S(i, \text{context})^{W_c}}{\sum_{k=1}^{N} S(k,j)^{W_e} S(k, \text{context})^{W_c}},$$

$$P_{\text{RECALL}}(i|j, \text{context}) = 1 - e^{-W_e S(i,j) - W_c S(i, \text{context})}.$$

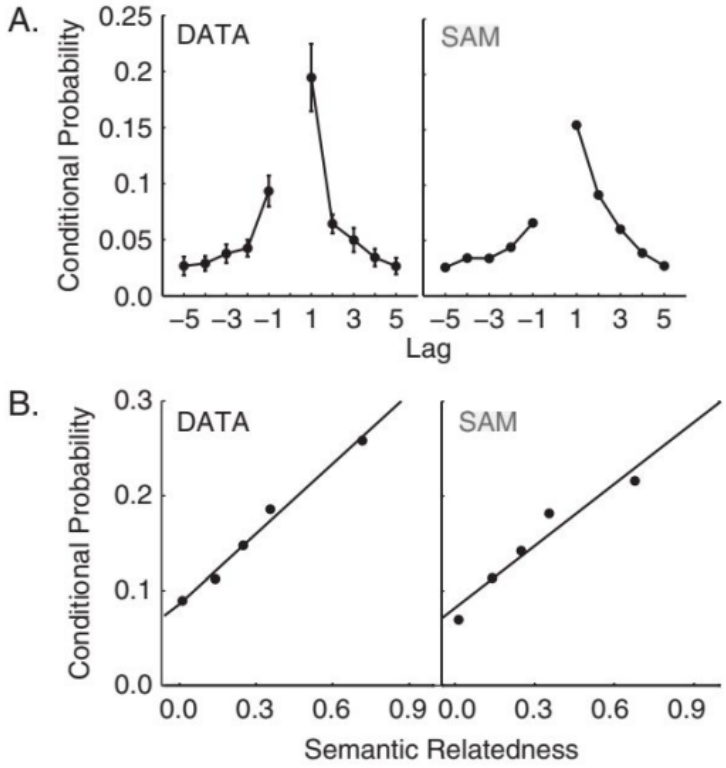
Here $S(i.j)$ is the strength of association in the LTS between items i and j, wwhile $S(k, context)$ is the strength of association between item k and the context. W_e and W_c are parameters that weight the relative importance of context and other list items. In the case of recalling the very first item, the context alone is used. Once recalled, an item will not be recalled a second time. When there have been Max consecutive failures at recall using a particular item as retrieval cues, SAM assumes that the simulated participant reverts to using the context alone as a retrieval cue.



The SAM model provides a good empirical match and an explanation for serial order effects in free recall of list items. Early list items are recalled more accurately because they spend more time in STS than later items (since these are kicked out more quickly once the STS fills up). More recent items are better recalled because they are likely to already by in the STS at the time of recall. This effect is eliminated by a distraction task, which eliminates the recency but not the primacy effect. Also the list length effect is explained because the probability of retrieving any given item during sampling decreases if there are more items competing.



SAM can also account for the tendency of subjects to recall items nearby in the list and items that are semantically related to each other around the same time, since they will have higher association strengths in the LTS (semantic association requires that the original SAM be augmented with a semantic-associative matrix, which aids in retrieval).



SAM is a process cognitive model, meaning that it attempts to capture the nature not just of the outputs that the recall process produces, but also the storage structures and the algorithm by which it is accomplishing the task. This stands in contrast to the probabilistic models you covered in the

first module of the class, where the emphasis is on identifying the information that are important to the task, while remaining agnostic about the algorithms that cognitive system is performing.

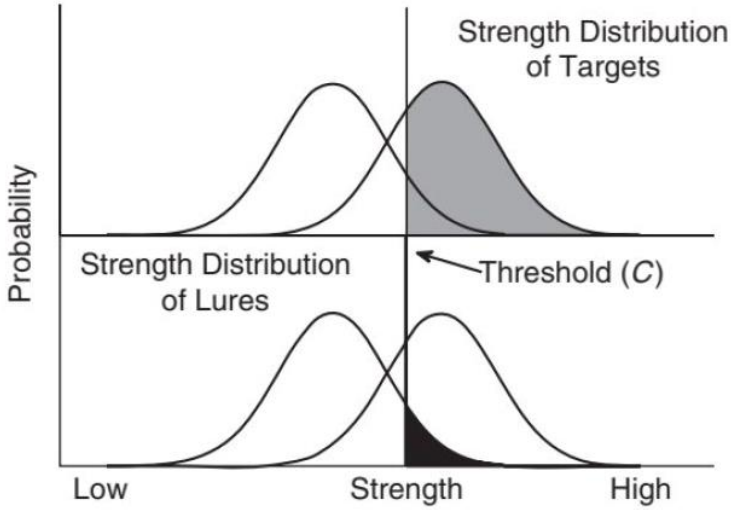
Signal Detection Theory

Signal detection theory describes the ability of subjects to discriminate between similar items. This is a form of recognition rather than free recall memory.

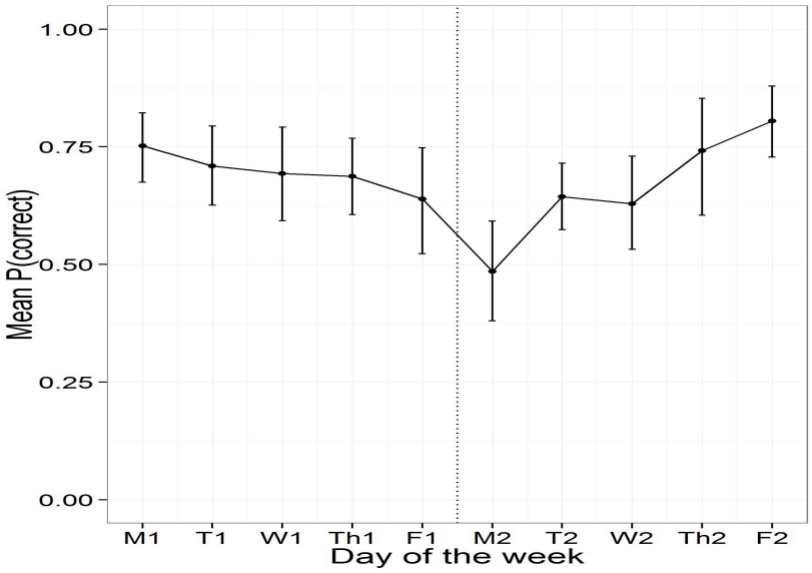
Classification of Responses in a yes–no Recognition Task.

	Response	
	"yes"	*"no"*
Target (Old)	Hit	Miss
Lure (New)	False Alarm	Correct Rejection

This can be modelled as selecting which of two Gaussian distributions has the largest density at the location of the signal. This method is called maximum a posterioi.



Signal detection theory predicts that differentiation is more difficult at the boundary between two classes, as the distributions overlap more in that location. This is consistent with data from various human experiments, such as shown below for subjects recalling which week an event occurred in.

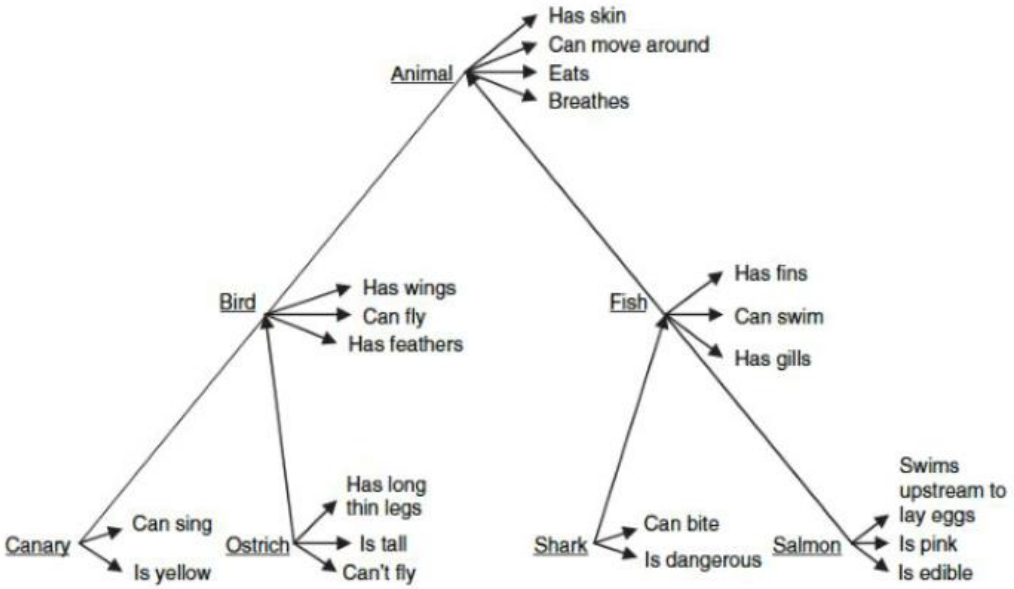

1.00
0.75
0.50
0.25
0.00
Mean P(correct)
M1
T1
W1
Th1
F1
M2
T2
W2
Th2
F2
Day of the week

Semantic Memory

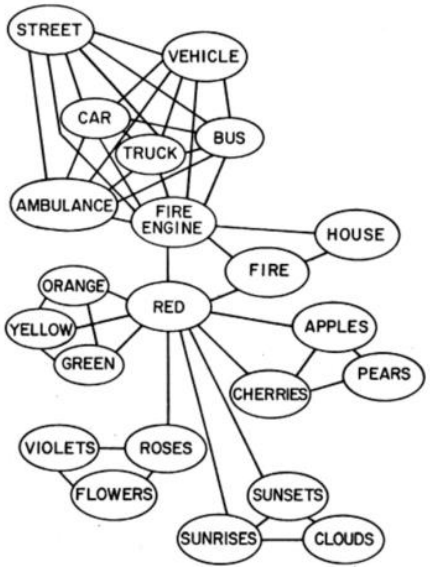
Semantic memory refers to general world knowledge of facts, ideas, meaning and concepts. It is distinct from episodic memory, which is our memory of experiences and specific events.

Feature-Based Methods

The semantic network was originally proposed as a hierarchical model of semantic memory in which concepts were nodes and propositions were labelled links (e.g., the nodes for dog and animal were connected via an “is a” link). The superordinate and subordinate structure of the links produced a hierarchical tree structure. Accessing knowledge required traversal of the tree to the critical branch, and the model was successful in this manner of explaining early sentence verification data from humans (e.g., the speed to verify that “a canary can sing” versus “a canary has skin”). However, this model is not good at explaining fast negative responses, which the model would predict should be slow as it would require complete traversal of the network.

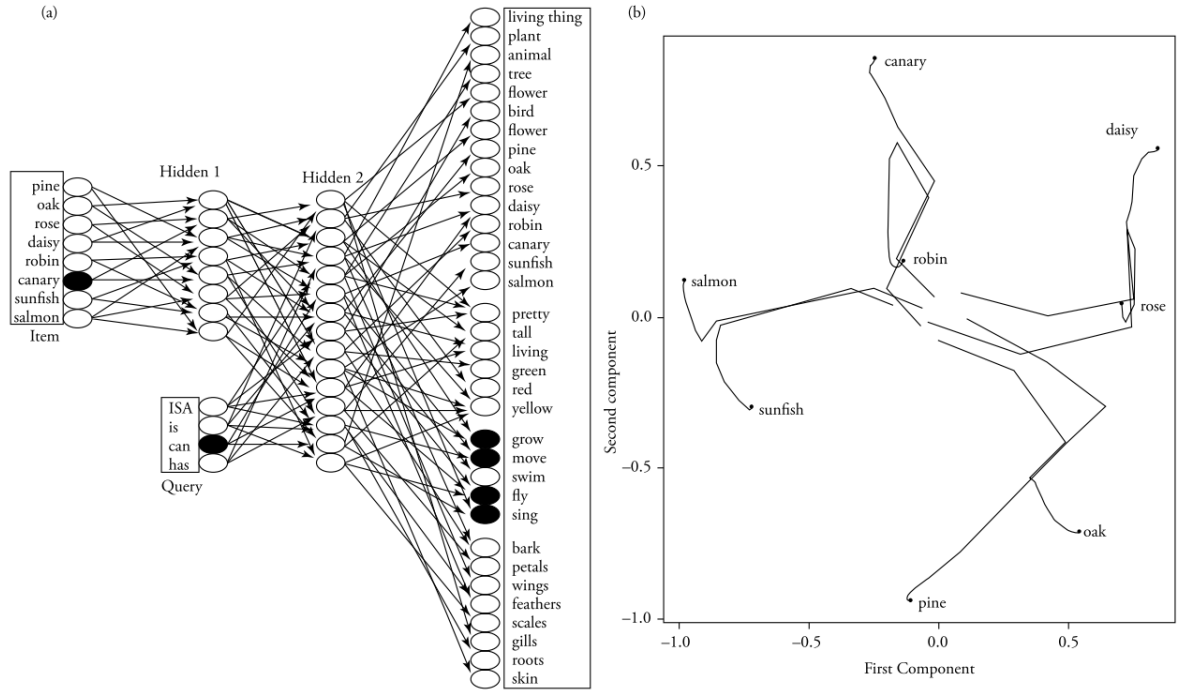


Other semantic network models emphasise the process of spreading activation through all network links simultaneously to account for semantic priming phenomena.



Connectionist Networks

In connectionist models, all concepts are represented across a common set of hidden units. When the network learns to associate 'robin+can' to 'sing' and 'fly', associations to other similar words like 'canary' will also be learned. The internal representations for the concepts show progressive differentiation, learning broader distinctions first and more fine-grained distinctions later. These models also provide a learning mechanism, which feature-based methods typically do not.



The general conclusion that Rogers and McClelland draw from this model is that a number of properties of the semantic system, such as the taxonomic structure and of causal knowledge, can be explained as an emergent consequence of simple learning mechanisms, and that these structural factors do not necessarily need to be explicitly built into models of semantic memory.

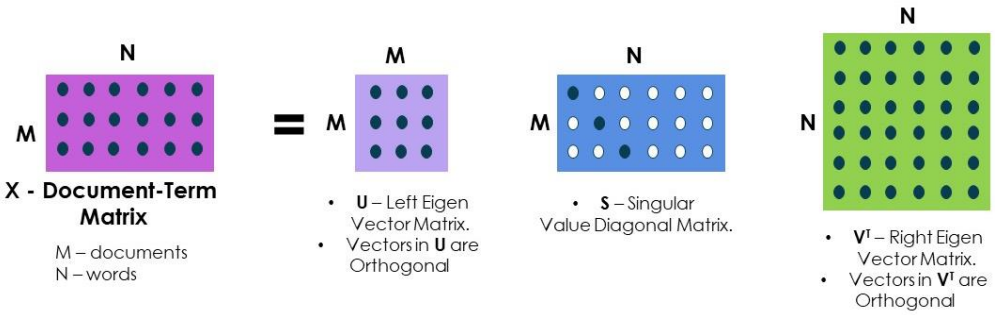
Distributional Models

There are now a large number of computational models in the literature that may be classified as distributional. Other terms commonly used to refer to these models are corpus-based, semantic-space, or co-occurrence models. The key idea of all such models is that words can be given an embedding in an underlying semantic space, which is derived from co-occurrence statistics of the words in question.

Latent Semantic Analysis (LSA) is a very popular technique which first constructs a term-by-document frequency matrix of a text corpus, in which each row vector is a word's frequency distribution over documents. A document is simply a "bag-of-words". The matrix is normalised by word frequency, then factorized using singular-value decomposition (SVD) into three component matrices: U, Σ, V. The U matrix represents the orthonormal basis for a space in which each word is a point, V represents an analogous orthonormal document space, and Σ is a diagonal matrix of singular values weighing dimensions in the space. More commonly, only the top N singular values of Σ are retained, where N is usually around 300. A word's semantic representation is then a pattern across the N latent semantic dimensions.

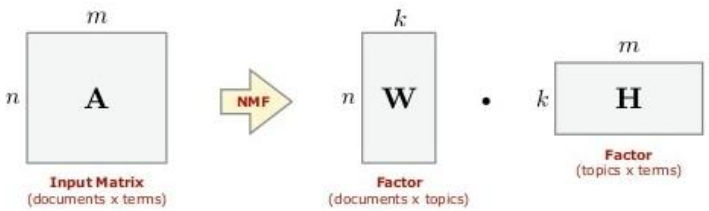
TABLE 2.2
The 10 x 9 Type-by-Document Matrix With Type Frequencies Corresponding to the Titles in Table 2.1

Types	*Documents*								
	M1	M2	M3	M4	M5	B1	B2	B3	B4
Bread	0	0	0	0	0	1	0	1	0
Composition	0	0	1	0	1	0	0	0	0
Demonstration	0	1	0	0	0	1	0	0	0
Dough	0	0	0	0	0	0	0	1	1
Drum	0	1	1	0	0	0	0	0	0
Ingredients	0	0	0	0	0	0	1	0	1
Music	1	0	0	1	1	0	0	0	0
Recipe	0	0	0	0	0	0	0	1	1
Rock	1	0	0	1	0	0	0	0	0
Roll	1	1	0	0	0	1	1	0	0

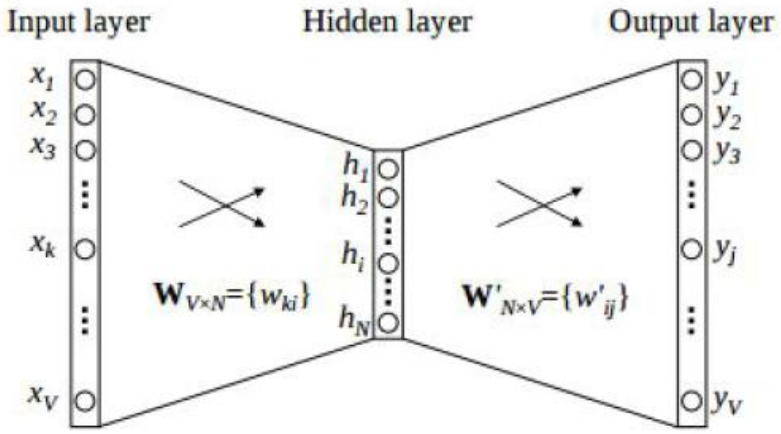


A variation of LSA are topic models. An assumption of a topic model is that documents are generated by mixtures of latent "topics," where a topic is a probability distribution over words. Although LSA makes a similar assumption that latent semantic components can be inferred from observable co-occurrences across documents, topic models go a step further, specifying a generative model for documents. The assumption is that when constructing documents, humans are sampling from a distribution of universal latent topics. To generate each word within this document, one samples a topic according to the document's mixture weights, and then samples words from that topic's probability distribution over words. To train the model, Bayesian inference is used to reverse the generative process: assuming that topic mixing is what generates documents, the task of the model is to invert the process and statistically infer the set of topics that were responsible for generating a given set of documents.

- Non-negative Matrix Factorization (NMF): Family of linear algebra algorithms for identifying the latent structure in data represented as a non-negative matrix (Lee & Seung, 1999).
- NMF can be applied for topic modeling, where the input is a document-term matrix, typically TF-IDF normalized.
- **Input:** Document-term matrix **A**; User-specified number of topics *k*.
- **Output:** Two *k*-dimensional factors **W** and **H** approximating **A**.



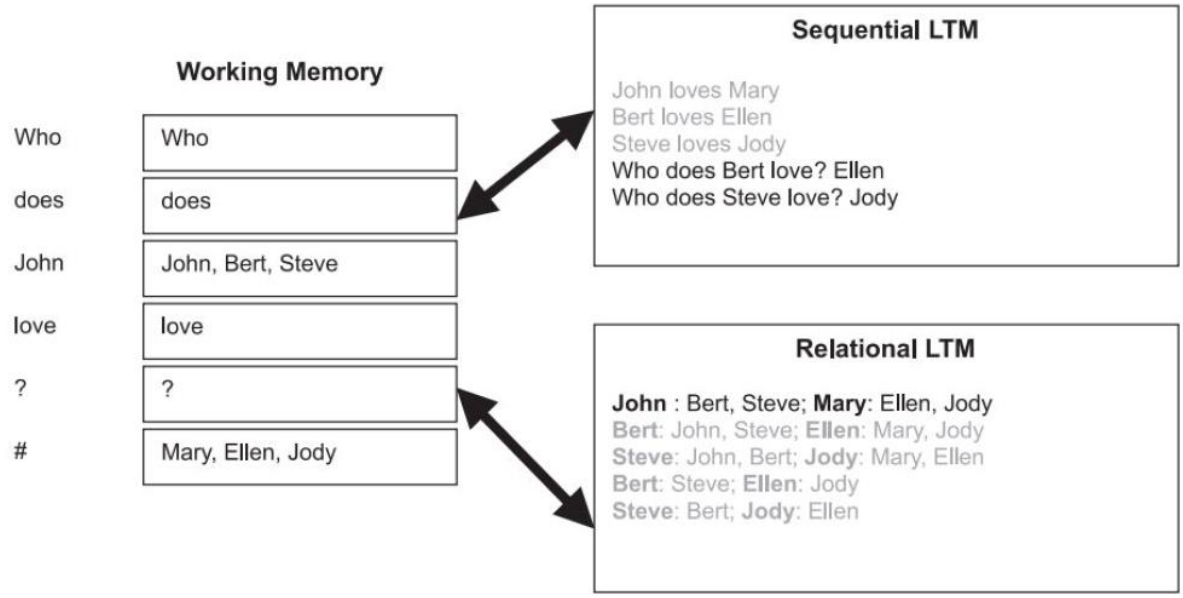
Other variants use repeated presentation of 'sliding window' segments of documents to a neural network, which learns compressed representations of particular words through the learned weights of the hidden units.



word2vec model architecture

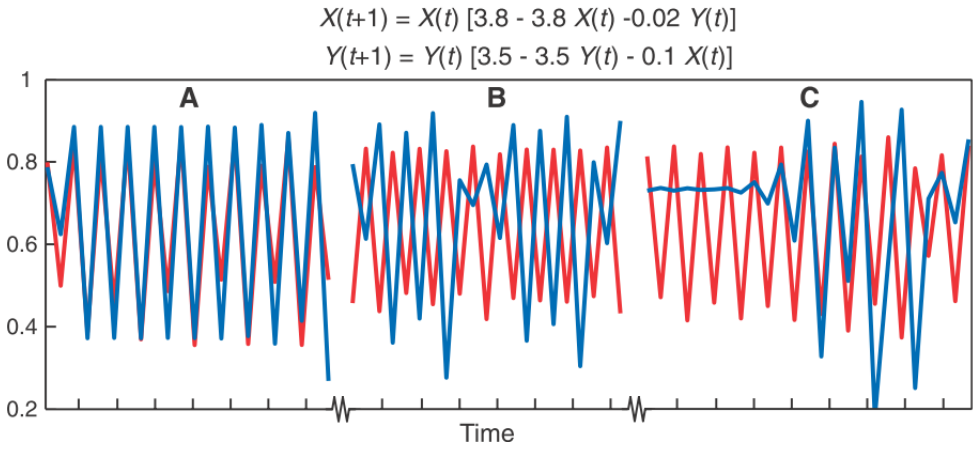
Compositional Semantics

Traditional distributional methods do not incorporate word order or role binding in embeddings. The study of how sentence structure determines role bindings is called compositional semantics. Dennis argued that extracting propositional structure from sentences revolves around the distinction between syntagmatic and paradigmatic associations. Syntagmatic associations occur between words that appear together in utterances (e.g., run fast). Paradigmatic associations occur between words that appear in similar contexts, but not necessarily in the same utterances (e.g., deep and shallow). The syntagmatic paradigmatic model proposes that syntagmatic associations are used to determine that words could have filled a particular slot within a sentence. The set of these words form role vectors that are then bound to fillers by paradigmatic associations to form a propositional representation of the sentence.



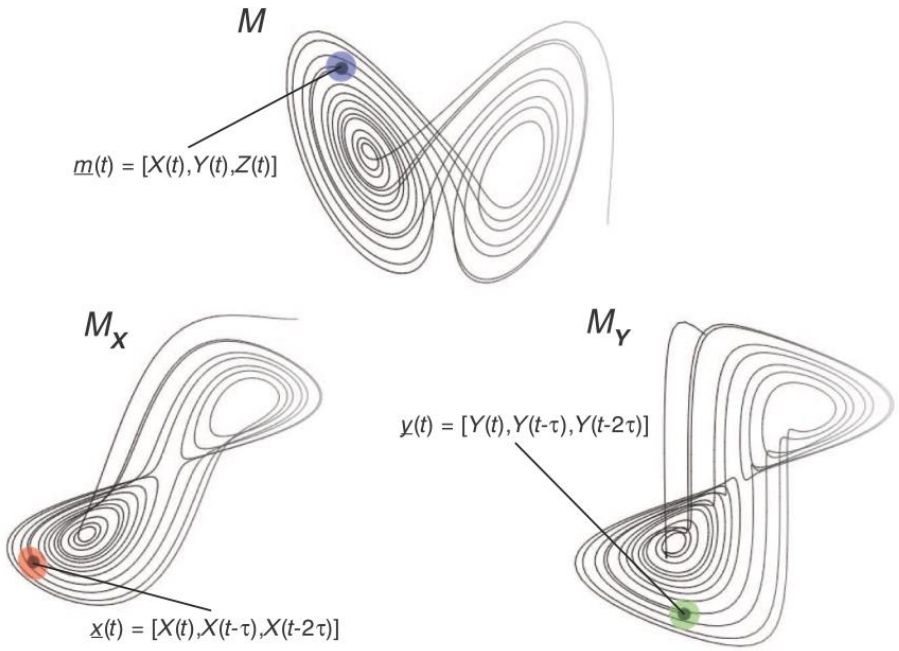
Convergent Cross Mapping

In nonlinear systems, variable correlation may not be at all related to whether the variables are causally connected. The following coupled equations show correlation at some times, anticorrelation at other times, and no correlation at all at yet other times. This highlights the need for better techniques to assess causal relationships in nonlinear systems.



An alternative approach called convergent cross mapping (CCM) instead tests for causation by measuring the extent to which the historical record of Y values can reliably estimate states of X. This happens only if X is causally influencing Y.

It works by first constructing the 'shadow manifold' for each variable of interest (in this case M_X and M_Y for X and Y respectively), from manifold M for the original system. This is done using lagged-coordinate embeddings of X and Y. Because X and Y are dynamically coupled, points that are nearby on M_X (e.g., within the red ellipse) will correspond temporally to points that are nearby on M_Y (e.g., within the green circle). This enables us to estimate states across manifolds using Y to estimate the state of X and vice versa using nearest neighbours.



One shadow manifold is used to predict the other by taking a set of E points on M_X, and then generating a series of predictions of Y (denoted $\hat{Y}|M_X$) using the equation:

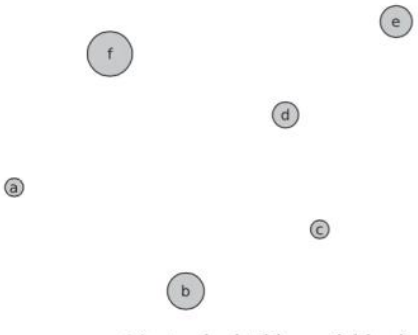
$$\hat{Y}|M_X = \sum_{i=0}^{E} w_i Y(t_i)$$

Here w_i is a weighting based on the distance between $x(t)$ and its ith nearest neighbor on M_X. Effectively, this amounts to assuming that Y is scattered about the manifold in the same pattern as X. If Y can be predicted from X with significant accuracy, we conclude that Y CCM causes X.

Social Network Models

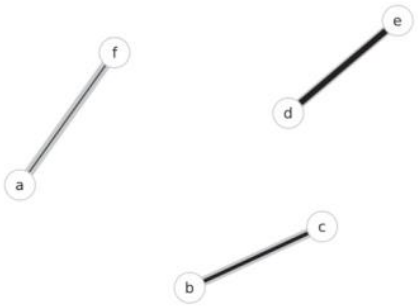
Introduction to Networks

Network science as the study of the collection, management, analysis, interpretation, and presentation of relational data. The roots of network science are particularly strong in social psychology, sociology, and anthropology. The key assumption of network science is that considering a set of nodes and ties as an interconnected unit adds analytical value to considering the nodes separately or even looking at dyads.



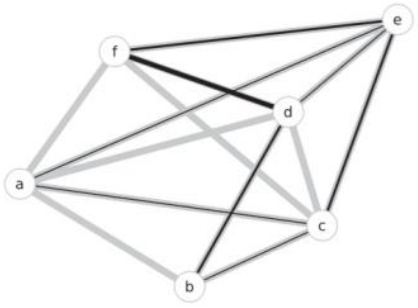
A	x
a	1
b	4
c	1
d	2
e	3
f	6

(a) standard table: variables in columns indexed with unrelated entities



D	x
(a,f)	1
(d,e)	5
(b,c)	3

(b) dyadic: variables in columns indexed with unrelated pairs of entities

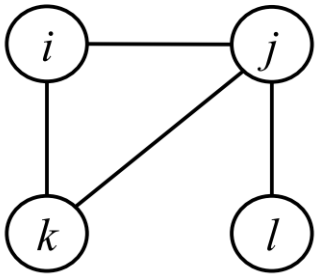


D	x
(a,b)	0
(a,c)	1
(a,d)	0
(a,e)	1
(a,f)	0
(b,c)	1
(b,d)	2
(c,d)	0
⋮	

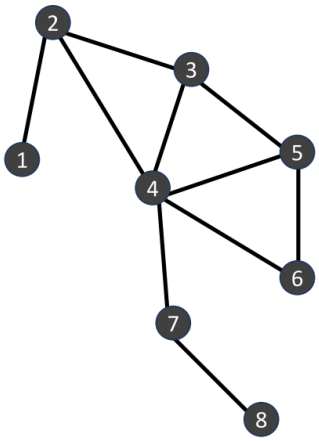
$x(D)$	a	b	c	d	e	f
a	·	0	1	0	1	0
b	0	·	1	2	·	·
c	1	1	·	0	2	0
d	0	2	0	·	1	4
e	1	·	2	1	·	2
f	0	·	0	4	2	·

(c) network: variables in columns indexed with incident pairs of entities, or in matrices

A graph $G(V,E)$ consists of a set of nodes/vertices $V = \{i,j,k,l\}$, and a set of edges/ties $E = \{\{i,j\},\{i,k\},\{k,j\},\{j,l\}\}$.



Networks can be represented as an adjacency matrix, which will always be square and symmetrical.



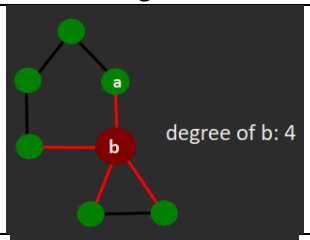
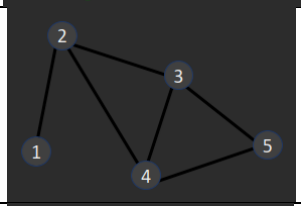
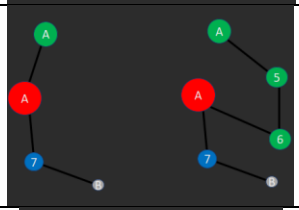
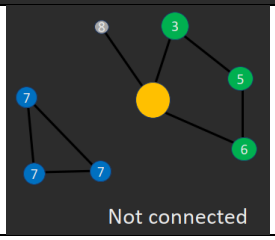
Set all NON-Ties to 0

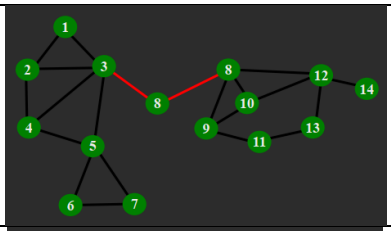
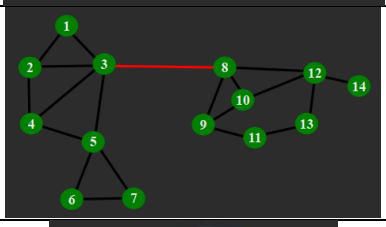
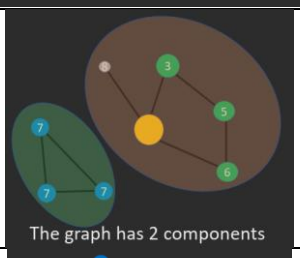
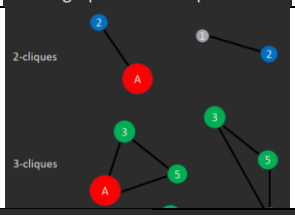
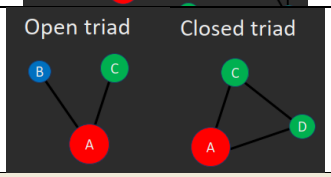
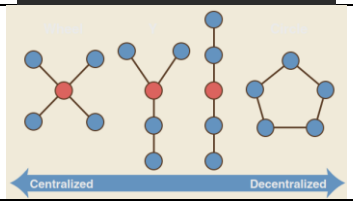
	1	2	3	4	5	6	7	8
1	-	1	0	0	0	0	0	0
2		-	1	1	0	0	0	0
3			-	1	1	0	0	0
4				-	1	1	1	0
5					-	1	0	0
6						-	0	0
7							-	1
8								-

Data for constructing network representations can be obtained through a variety of methods, including ethnography, archives, surveys, online databases, etc. A common method is called sociometric free recall, which involves asking participants to name a certain number of contacts (such as friends or workmates).

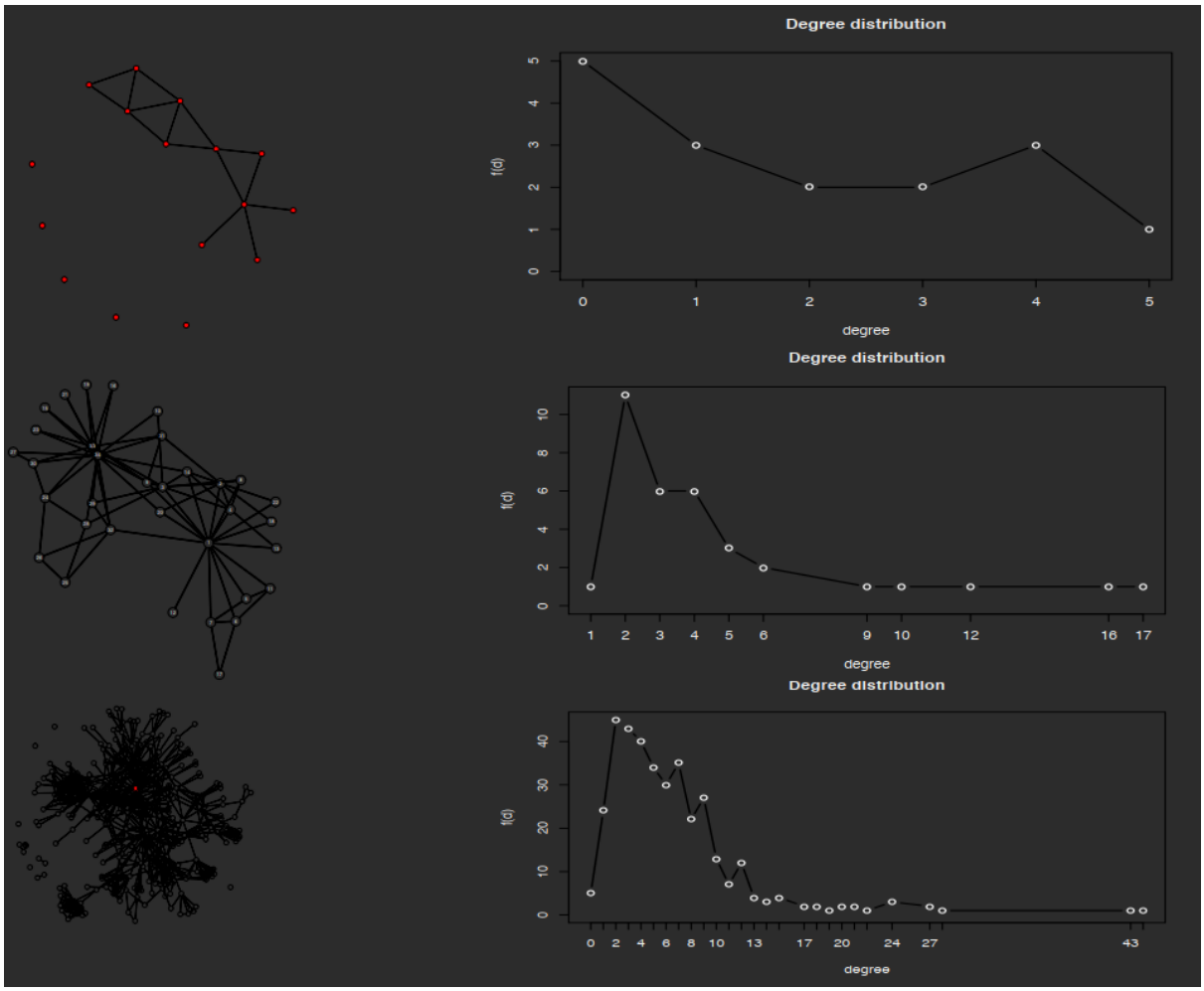
Network Characteristics

Table of key network terminology.

Concept	Formula	Explanation	Figure
Node degree	$d_i = \sum_j X_{ij}$	The number of ties of a node.	
Graph density	$d(G) = \frac{\sum_{i<j} X_{ij}}{n(n-1)/2}$	The total number of ties divided by the number of possible ties.	
Geodesic distance	$\min_k \sum_k X_{ik} X_{kj}$	The length of the shortest path between two nodes.	
Connectedness		A graph is connected if there is a path between any two nodes.	

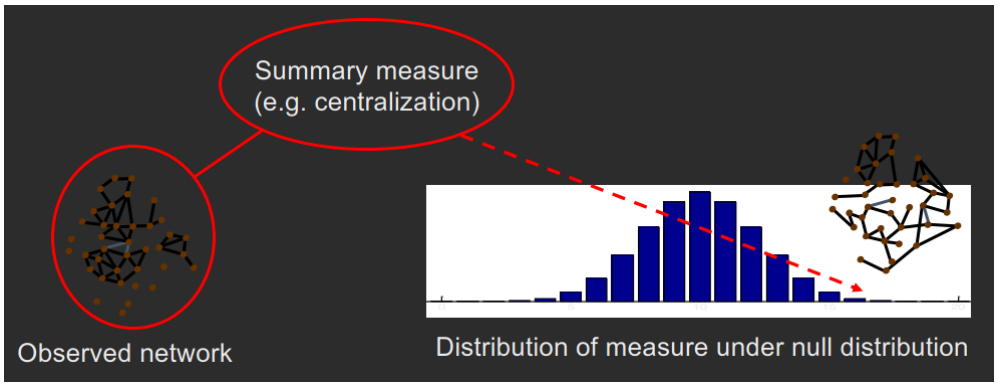
Cutpoint (node)		A node that connects the network.	
Bridge (edge)		An edge that connects the network.	
Component		A subgraph that is maximally connected.	
k-clique	$C_k = \{V_i\}$: $\lvert C_k \rvert = k$, $\sum E_{ij} = k(k-1)/2$	A subset of k nodes that are all connected. Cliques may be overlapping.	
Triad	T_3: closed triad T_2: open triads	A set of three nodes. A closed triad is a 3-clique.	
Centralisation index	$\sum_i \lvert d_{max} - d_i \rvert$ $d_{max} = \max_i(d_i)$	Measurement of how centralised ties are about few nodes.	
Clustering coefficient	$\frac{3T_3}{3T_3 + T_2}$	Measurement of how clustered triads are.	

The degree distribution is an important way to describe how connections are distributed across the network.



Random Graphs

One common strategy in network science is to compare observed networks to randomly generated networks that are in some sense 'similar' to the observed network, and then analyse the nature of the differences between the two. For example, in real networks people do not form ties at random, so we can compare how the real network differs from a network where ties are randomly formed.



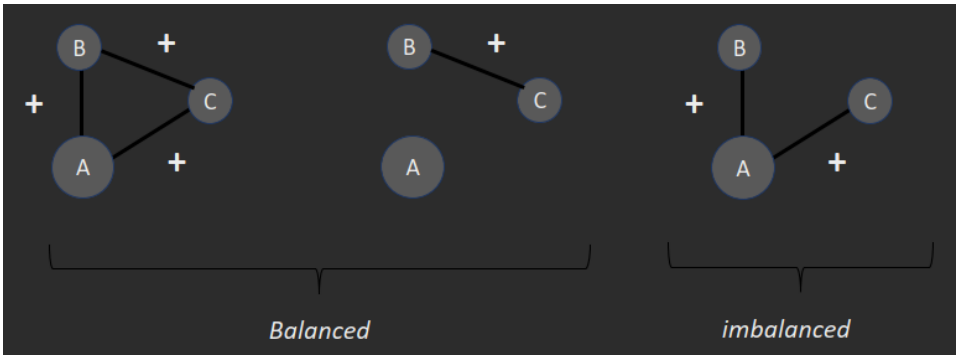
Assume that there is no mechanism A
Let $T(X)$ be a statistic that is sensitive to A (e.g. larger the stronger A is)
Let $p(X)$ be a distribution that does not have A
Test
H_0: tie-formation is not driven by A, against
H_1: tie-formation is driven by A
If $X \sim p(X)$, H_0 is true, we can calculate the probability
$\Pr(T(X) \geq k)$
Reject H_0 on the α-level if
$\Pr(T(X) \geq T(X_{obs})) < \alpha$

Random Graph Models Table

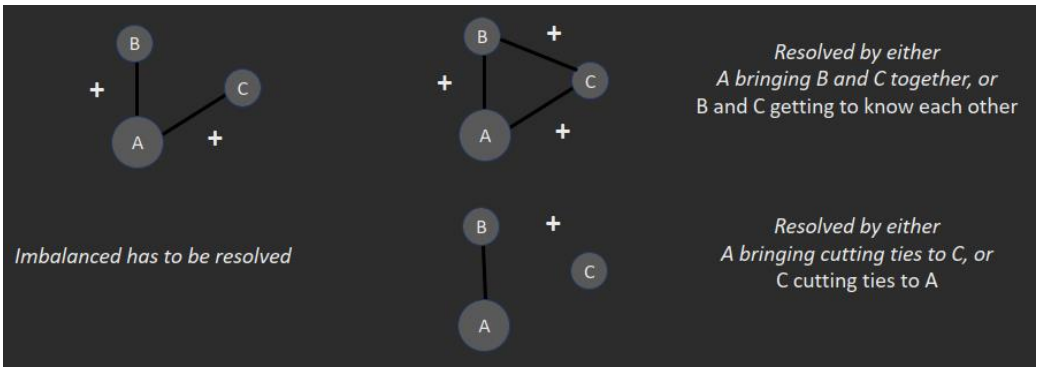
Random graph name	Formula	Description
Bernoulli random graph (BRG)	$X_{ij} \sim Bern(p)$ $d_i = \sum_j X_{ij} \sim Bi(n-1, p)$	Actors form ties completely at random with probability equal to network density. All nodes have same degree distribution.
Conditionally uniform density (U|L)	$X \sim U\|L(X) = k$	Network has a uniform distribution of ties, conditional on the density. Produced by randomly shuffling the adjacency matrix.
Conditionally uniform degrees (U|d)	$X \sim U\|d(X_i)$	Network has a uniform distribution of ties, conditional on the degree distribution for each node. Produced by randomly shuffling each row of the adjacency matrix.
Exponential Random Graph Model (ERGM)	$P(X) = \exp\left[\sum_s \theta_s z_s(X) - \phi(\theta_s)\right]$ $z_s(X)$: number of feature s θ_s: weight for feature s $\phi(\theta_s)$: normalising constant	A general framework where probability of a network is given as a weighted sum of features z_s. Features can be number of edges, two-stars, triangles, etc.
Directed networks: Conditional uniform (MAN)	$X \sim U\|MAN(X)$	Network has a uniform distribution of ties, conditional on the dyad census. Number of mutual (M), asymmetric (A), and null (N) dyads is fixed.
Directed networks: Conditionally uniform degrees (U|din,dout)	$X \sim U\|d_{in}(X_i), d_{out}(X_i)$	Network has a uniform distribution of ties, conditional on the in and out degrees for each node.

Balance Theory

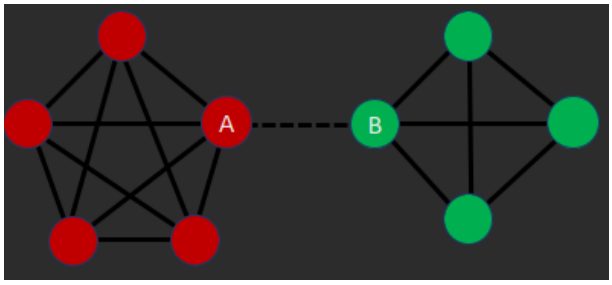
Balance theory predicts that only particular types of network structure will be stable. In particular, it says that triads with mismatching connections will not be stable and hence are rarely found in social networks.



Imbalances may be resolved by two different means:



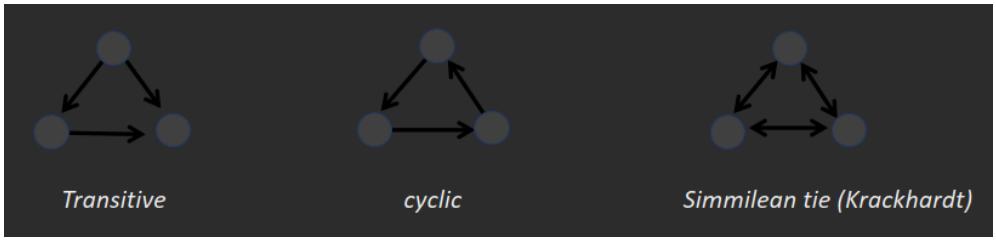
The result of such a process will be that strong ties will tend to cluster into cliques. However, weak ties do not have the same balance requirements, and so The global structure will tend to be of cliques of strong ties, connected by weak ties.



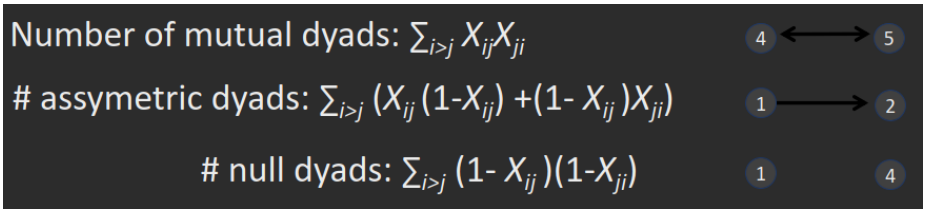
It has been hypothesised that these weak ties are especially important, as they enable ideas or goods to flow between cliques across long distances of the network.

Directed Networks

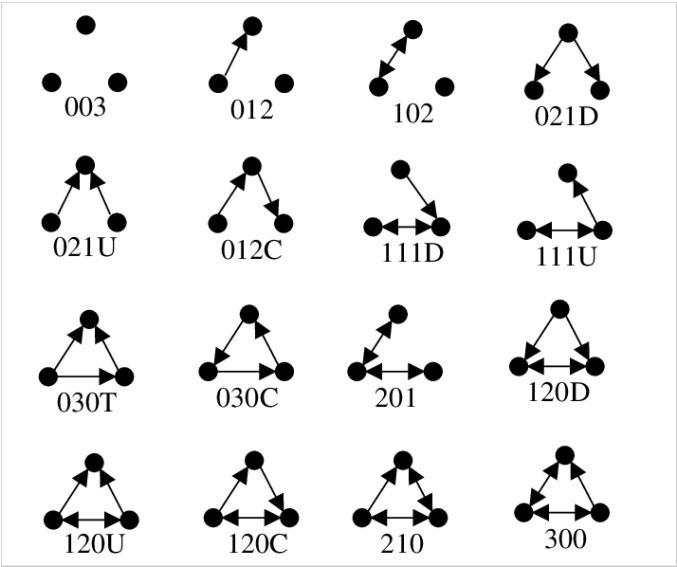
Directed networks allow there to be an asymmetry between ‘incoming’ and ‘outgoing’ ties from one node to another. These can represent different things in different contexts. IN some cases we might expect ties to be bidirectional, while in other cases we might expect intransitive ties (e.g. supply chains), and in others unidirectional ties (e.g. mentorship).



In directed networks, there are three different types of dyads.



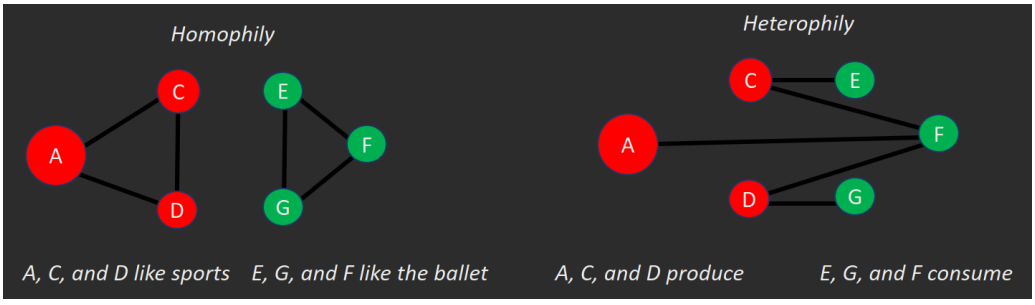
There are sixteen possible types of triads in a directed network. The distribution of the frequency of all types is called a triad census.



Network Contagion and Influence

Homophily

Homophily is the principle that a contact between similar people occurs at a higher rate than among dissimilar people. The pervasive fact of homophily means that cultural, behavioral, genetic, or material information that flows through networks will tend to be localized. While many networks show homophily, heterophily is also observed in certain cases, such as customer/supplier relationships.



Homophily of trait y can be measured as the sum of all tied nodes that have the same value of the trait (if y is binary). If the trait is a scaled value, then this measures the extent to which high values of y co-occur in tied nodes.

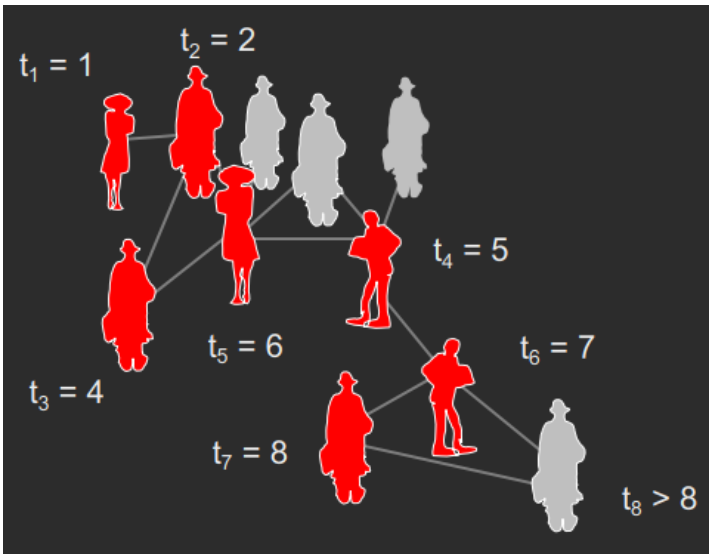
$$H(X) = \sum_{ij} y_i y_j X_{ij}$$

Homophily in networks can be explained by three main different mechanisms:

1. Social influence: people who are linked become more like each other.
2. Propinquity: people affiliated with similar things are similar.
3. Social selection: people who are similar select out other people like them to form ties with.

Event-history Analysis

Social influence can be investigated by studying time series data of networks. One question that can be considered (here in the case of innovation adoption) is whether the time to adoption is shorter for individuals that have a tie to someone who has already adopted compared to people that do not have a tie to someone who has adopted?



We can model the spread of the innovation (or whatever else we are interested in) as an Exponential distribution, where the time to adoption t of individual i is distributed as:

$$f(t_i) = \lambda_i e^{-\lambda_i t_i}$$

We want to know whether this time changes with the network ties the person has. We can test this using the following model:

$$t_1 - t_0 \sim \text{Exp}[\lambda_i(Y(t_0), X]$$

Where $Y_i(t)$ is an indicator for whether individual i has adopted the innovation by time t, X is the network, and λ_i is the rate of adoption for individual i. We can model the adoption rate as a function of the number of ties the individual has with other adopters as follows:

$$\lambda_i\big(Y(t), X(t)\big) = \exp\big[\alpha + \beta a_i\big(Y(t), X(t)\big)\big]$$

There are many possible choices for the influence dependency function a_i. Some common choices are shown in the table below.

Equation name	Formula	Explanation
Total exposure	$a_i(y,x) = \sum_j y_j x_{ij}$	The more people you know that have adopted, the quicker you will adopt.
Weighted total exposure	$a_i(y,x) = \sum_j y_j x_{ij} s_j(x)$	As above, but with importance weighting $s(x)$ for each tie.
Average exposure	$a_i(y,x) = \frac{\sum_j y_j x_{ij}}{\sum_j x_{ij}}$	The greater the proportion of people you know that have adopted, the quicker you will adopt.

The adoption times for the entire network $t_1, t_2, \ldots, t_k$ are descripted by the likelihood function, which is the product of the individual pdfs:

$$\prod_{i=1}^{k} \lambda_i\big(Y(t_i), X(t_i)\big) \exp\left[-\sum_{i=1}^{k} \lambda_i\big(Y(t_i), X(t_i)\big)(t_i - t_{i-1})\right] \times \prod_{i=1}^{k} \big(1 - \exp\big[-\lambda_i\big(Y(t_i), X(t_i)\big)(t_i - t_{i-1})\big]\big)$$

Parameters can be estimated using Maximum likelihood or Bayes.

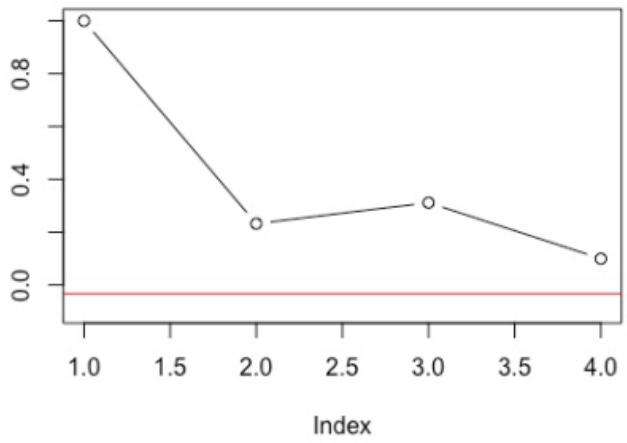
Auto-correlation Models

When estimating network regression models, it is important to pay close attention to the residual terms. Consider a simple regression model:

$$Y_i = \alpha + \beta M_i + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

Where M_i is some variable of interest for individual i in the network X.

The problem with this simple model is that error terms are unlikely to be independent, as we have ignored the effects of the network. Tied individual are more likely to be similar in various ways, and so probably will have correlation error terms. This can be tested by examining the autocorrelation figure (past index 1, which is just the node itself):



To deal with this correlation, we can instead estimate an autocorrelation model:

$$Y_i = \alpha + \beta M_i + \epsilon_i$$
$$\epsilon_i = \rho \sum_j X_{ij}\epsilon_j + \xi_i$$
$$\xi_i \sim N(0, \sigma^2)$$

One problem with this formulation is that more connected nodes have higher errors. To deal with this we can scale by the total number of ties:

$$Y_i = \alpha + \beta M_i + \epsilon_i$$
$$\epsilon_i = \frac{\rho \sum_j X_{ij}\epsilon_j}{\sum_j X_{ij}} + \xi_i$$
$$\xi_i \sim N(0, \sigma^2)$$

An alternative to modelling interaction through the error terms, we can model it directly through the main part of a more complex social influence model:

$$Y_i = \rho \sum_j X_{ij}\epsilon_j + \alpha + \beta M_i + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$

Network Empirical Techniques

Negative Ties

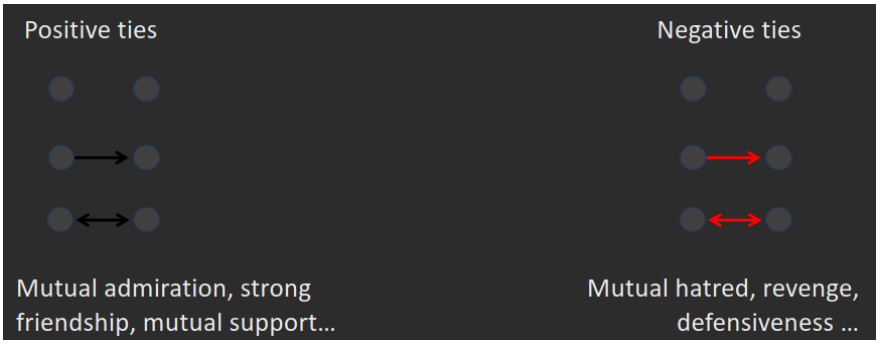
A negative tie is a connection which conveys something detrimental, such as harm or hostility. They are often rarer than positive ties, and have been less often studied, but can be very important. Networks with both positive and negative ties are called signed networks. Dynamics of negative and signed ties are qualitatively different. For example, while we might expect positive ties to be protective against reciprocation with negative ties, the opposite is often true; esteem ties tend to be reciprocated with disesteem ties.

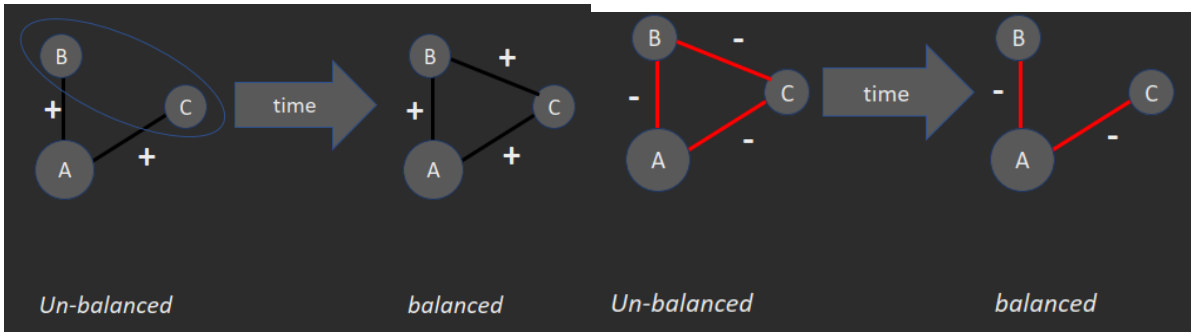
Table 1
Examples of potential negative ties across a range of social network settings.

Setting	Actors	Potential Negative Ties
1. Family	Family members	Conflict, dislike, avoidance, domestic violence, arguments, defensiveness, silence, contempt, tease, bully
2. Workplace	Employees, staff	Negative gossip, discipline, bullying, avoidance, arguments, exploitation, dismissal/termination, judge incompetent, sabotage, undermine legitimate interests
3. School	Students	Bullying, avoidance, dislike, conflict, arguments
4. Romantic and sexual relations	Persons	Reject, find unattractive, avoid, break-up, argue, violence, give STD, harrassment, assault, rape
5. Friendship and socialising	Persons	Conflict, argue, dislike, avoid, hit, tease
6. Social Media	Persons, accounts	Argue against, negative retweet, express dislike, distain
7. Internet	Websites, Wikipedia accounts	Delete/reverse edits, vote against, critical review, low rating
8. Neighbourhoods	Persons, households	Argue, conflict, dislike, avoid, legal or police complaint, robbery or theft
9. International relations	Nation states	Public criticism, vote against in UN committees (or similar), sanctions, threats of aggression, aggression (war)
10. Economic relations	Nation states	Sanctions, tariffs, currency manipulation, boycotts, bans, quotas, non-tariff barriers
11. Politics	Parliamentarians, Ministers, Parties, lobby groups	Vote against, public criticism, negative advertisements, conflicting policy positions
12. Voluntary organisations	Organisations	Public criticism, campaigns against
13. Inter-firm	Firms	Legal suit, public criticism, competing products, shared market
14. International organisations	Organisations	Public criticism, conflicting policy positions, vote against

Assessing negative ties through data can be very difficult, because of different social perceptions. For example, Tatum and Grund (2019) find there is very little corroboration between perceptions of whom a teen student feels they are bullying and those presumed victims expressing that they are the targets of that individual's behaviours. Also, it is notoriously difficult to get people to provide negative nominations.

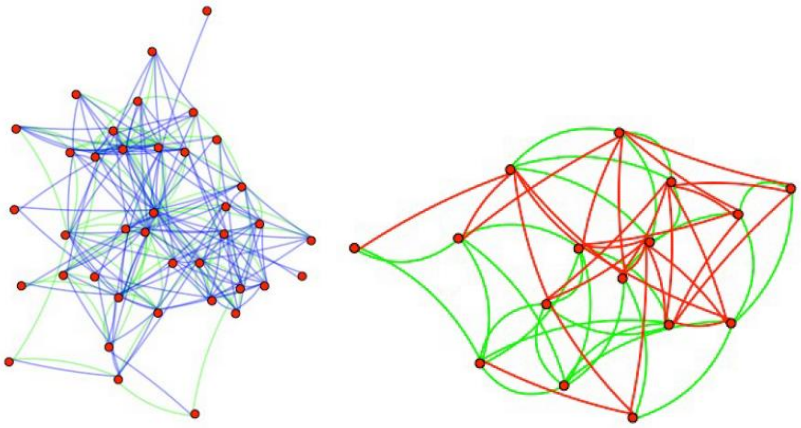
Much research indicates that negative ties have pervasive effects throughout the network, as people tend to avoid negative ties and hence they hinder the function of the entire network. The dynamics of negative ties is also different than for positive ties, since fully connected triads are unbalanced, and negative ties are less often reciprocated, since avoidance is typically easier.



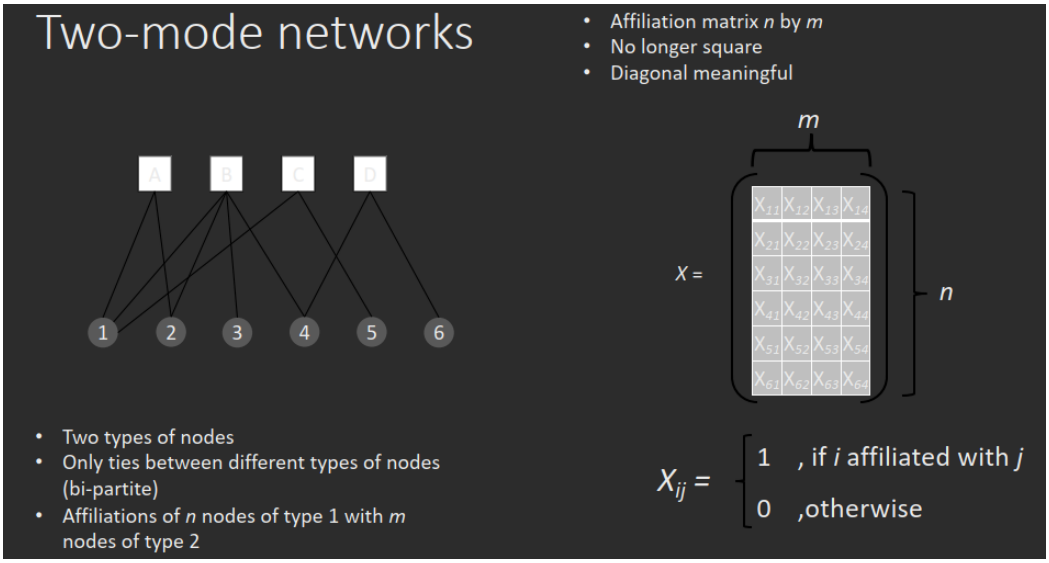


Multiplex Ties and Nodes

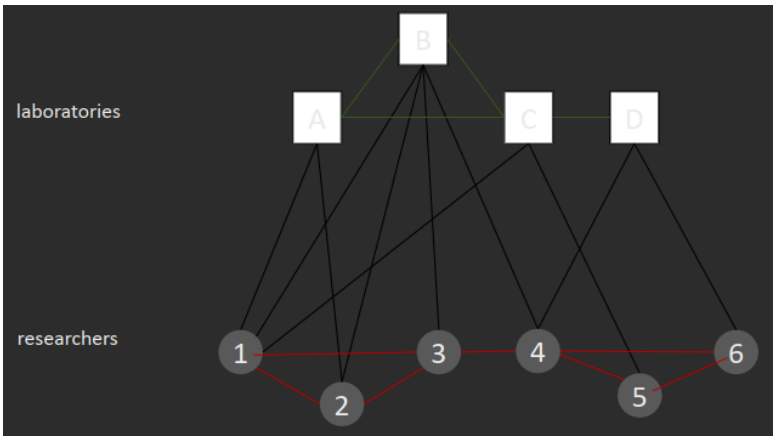
Some graphs include multiple different types of ties. These could be positive and negative, or simply different types of connections, like co-workers and friends.



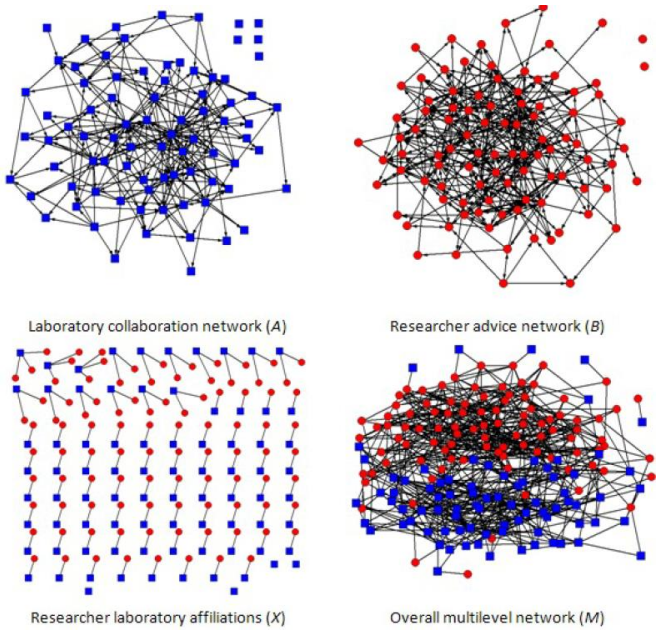
Networks can also have different types of nodes, such as individuals and organisations, or buyers and sellers. Two-mode networks are a special case of this, where there are two types of nodes, and ties are only allowed between nodes of different types. Adjacency matrices are no longer square.



This can be extended by combining two-mode networks with one-mode networks within similar types of nodes (compare black ties with red ties).

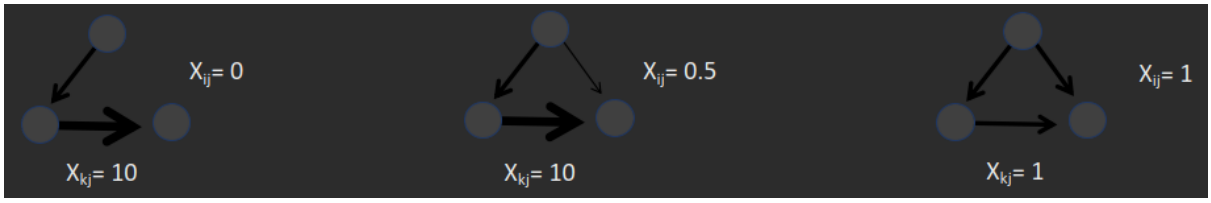


One study, for example, used this approach to study researcher advice and laboratory collaboration relationships for French cancer researchers.



Valued Networks

In valued networks, ties are no longer binary, but can take a range of different values. This makes it difficult to apply many of the existing techniques and concepts for traditional graphs. For example, in the triads shown below, which is more transitive? It is unclear how to think about this.



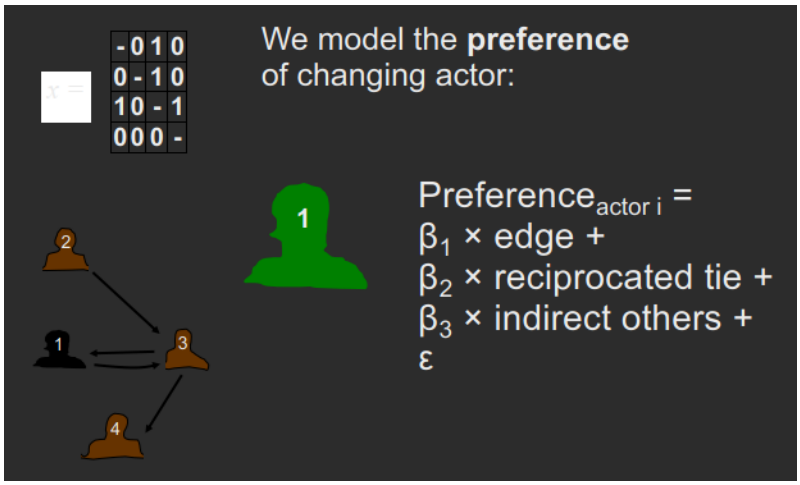
Actor-oriented Models

A dynamic network consists of ties between actors that change over time. These models are constructed using longitudinal data and assume that actors in the network actively take steps to change ties and shape the ties in the network. The changing network can be interpreted as the outcome of a Markov process, i.e., that for any point in time, the current state of the network determines probabilistically its further evolution. The model has directed ties, where each tie has a sender, who controls the sending of the tie, and a receiver.

The model works by assuming that at each point in time, every tie t_i has the probability of changing that is distributed according to the state of the network X:

$$t_i \sim \exp[\lambda_i(X)]$$

The function λ will typically include a range of effects, which are functions of the network that are chosen based on theory and subject-matter knowledge.



The results of such models can be used to determine the magnitude of social selection versus social influence effects.