Theoretical note: The relation between structure and dynamics in psychological
networks of attitudes

Mark G. Orr[1], Emily S. Teti[2], Andrei Bura[3] and Henning Mortveit[3]

[1]Florida Institute for Human and Machine Cognition

[2]Los Alamos National Laboratory, Nuclear Engineering and Nonproliferation Division

[3]University of Virginia, Biocomplexity Institute

Abstract

Two claims of the the Causal Attitude Network (CAN) model and the descendent Attitude Entropy framework (AE) are indicative of significant theoretical hurdles facing the psychological network modeling efforts of attitudes. The first claim is that the dynamics of change in an Ising-like attitude network, under perturbation of any one single node, can be inferred from the static network attributes of said node. The second claim is that psychological network models of attitudes with Ising-like dynamics will maximize both attitudinal consistency and accuracy when within the small-world topological regime. The first claim, one with significant application potentials, has not been sufficiently tested; the second claim, one with high theoretical novelty, has never been addressed. Using a set of analytic results and simulations, we found little support for these claims–in short, the predictions are not logically consistent with the theory. Our results have implications beyond attitude models to the the larger field of psychological networks (e.g., in clinical psychology) in reference to how we should explain and understand their dynamics. KEYWORDS: attitudes, neural networks, dynamical systems, psychological networks

Theoretical note: The relation between structure and dynamics in psychological networks of attitudes

Introduction

Over the past 25 years, the social psychological literature has successfully integrated some of the computational modeling approaches from cognitive science to address a range of phenomena: causal attribution, stereotypes, attitude formation, impression formation, and personality (e.g. Conrey & Smith, 2007; Monroe & Read, 2008; M. G. Orr, Thrush, & Plaut, 2013; F. V. Overwalle, 2007; Read & Miller, 1998; Smith, 1996; Vallacher, Read, & Nowak, 2017). Constraint satisfaction, typically formalized and implemented as a kind of recurrent artificial neural network, holds a strong position in social psychology. Not only does it fit past and current understandings of a range of phenomena, it offers a formalization of mechanism (see Read, Vanman, & Miller, 1997; Simon & Holyoak, 2002). Computational models of attitude formation and change, the topic of this article, are dominated by the constraint satisfaction formalism (e.g. Conrey & Smith, 2007; Ehret, Monroe, & Read, 2015; Monroe & Read, 2008; M. G. Orr & Plaut, 2014; M. G. Orr et al., 2013; F. Overwalle & Siebler, n.d.; Van Overwalle & Siebler, 2005).

Recently, a new class of attitude model has come into the fray. The Causal Attitude Network (**CAN**) model and close-cousin Attitudinal Entropy (**AE**) framework, have been cast, in their union, as a novel theoretical approach for understanding attitude formation and change (Dalege et al., 2016; Dalege, Borsboom, van Harreveld, & van der Maas, 2018; Dalege & van der Maas, 2020)[1]. This approach stems from the psychological networks approach that rose to prominence in the clinical psychology literature over the past decade or so (Borsboom, 2008, 2017; Borsboom & Cramer, 2013; Bringmann, 2021; Bringmann & Eronen, 2018; A. O. J. Cramer, Waldorp, Maas, & Borsboom, 2010). For purposes of this article, we will dub this class of attitude model as the CANAE approach or framework or, for simplicity, just CANAE.

The novelty of CANAE, on the surface, stems from: (i) its use of constructs from statistical physics (e.g., Gibbs and Boltzmann distributions of the configuration space, pseudo-thermodynamic temperature effects), (ii) its use of constructs from network science (e.g., global and local topological properties of the networks)[2], and (iii) its learning mechanism.

A bit deeper, however, we see another story. CANAE is an Ising-like model (Dalege et al., 2016) and, thus, falls within the larger class of constraint satisfaction models of attitudes. Others have used statistical physics counter-parts in constraint-satisfaction models of attitude, e.g., the use of energy in Monroe and Read (2008) or the use of attractors in M. G. Orr and Plaut (2014). Prior work has considered attitudinal stability as a property of the network topology, expressed in the form of the magnitude of the edges on a graph (Monroe & Read, 2008) or the network structure itself (M. G. Orr et al., 2013; Shultz & Lepper, 1996). In principle, then, the CANAE approach is very closely aligned with prior constraint satisfaction computational models of attitudes networks.

We claim that the real novelty of the CANAE framework lies in its use of constructs from network science and statistical physics to develop new kinds of

---

[1] It is referred to as "new."(Dalege et al., 2016, see p. 3, col 1, last paragraph)

[2] Under the umbrella of *psychological networks*

theoretical predictions, e.g., in respect to the relation between the structure of the attitude network to its dynamics. The natural path in such theoretical enterprises, to begin empirical testing of said predictions, has commenced (Chambon, Dalege, Elberse, & van Harreveld, 2022; Dalege, Borsboom, van Harreveld, & van der Maas, 2017, 2019; Dalege, Borsboom, van Harreveld, Waldorp, & van der Maas, 2017; Dalege & van der Does, 2022; Zwicker, Nohlen, Dalege, Gruter, & van Harreveld, 2020).

We think some of these empirical efforts were premature. Some of the new theoretical predictions of CANAE have not been verified to be logically consistent with the model–i.e., they may not be derivable from the CANAE model.

Our main objective in this article is to verify the logical consistency of two foundational CANAE claims. The first claim is that understanding the local topological characteristics of individual components in an attitude network is sufficient to predict how perturbation of said components (e.g., via persuasion) will affect the global dynamics of attitude formation. This issue is of practical import for clinical psychology (see Bringmann, 2021; Bringmann et al., 2019; Bringmann & Eronen, 2018; Burger et al., 2020; A. O. Cramer et al., 2016; Haslbeck, Ryan, Robinaugh, Waldorp, & Borsboom, 2021; Wichers, Wigman, & Myin-Germeys, 2015) and other applied domains in which persuasion or behavior change are paramount for prevention, intervention and mitigation (e.g., public health (M. Orr, Mortveit, Lebiere, & Pirolli, 2023; M. G. Orr & Chen, 2017; M. G. Orr & Plaut, 2014; M. G. Orr, Zeimer, & Chen, 2017), climate change (Thompson, 2023), disaster preparedness (Schlegelmilch & Carlin, 2023)). The second claim is that the structure of attitude networks is constrained by a drive to optimize a trade-off between attitudinal accuracy and attitudinal consistency. (The desired state is relatively high-consistency without losing too much accuracy.) The network structure that affords such a trade-off is small-world (see Watts & Strogatz, 1998, for the small-world algorithm). This claim does not reflect immediate practical concerns, yet, its theoretical novelty and subsidiary implications for learning in CANAE are motivation enough to warrant verification.

In the remainder of this article we will: (i) provide the necessary technical background on CANAE; (ii) offer two in-depth studies of the two target CANAE claims; and (iii) conclude with discussion and conclusions.

But first, we want to frame our work in reference to one of the central scientific issues in contemporary scientific psychology: the so-called replication crisis (see Nosek et al., 2022). Our work exemplifies an unsung heroine of the replication crisis: formal computational and mathematical modeling. In contrast to the oft-sung heroes of this crisis–better data, better statistical methods and deeper administrative controls (e.g., pre-registration)–computational and mathematical modeling are of limited repute, largely due to the degree to which they are misunderstood in terms of use and value. A small, nascent effort to mend these rifts has erected a motley set of arguments for the necessity of formal modeling efforts in scientific psychology (e.g., Fried, 2020; Oberauer & Lewandowsky, 2019; Robinaugh, Haslbeck, Ryan, Fried, & Waldorp, 2019; Smaldino, 2020), something that is aimed, by necessity, more towards social and clinical psychology. The cognitive sciences, neural sciences and perceptual sciences use and train with such models with more regularity. Note, however, that there is an argument for all of scientific psychology, including these latter sub-disciplines, to persist in driving towards the development of a mature science, one with a cumulative, systematic march to overarching theoretical clarity (see Muthukrishna & Henrich, 2019). Our work presented here hopes to contribute towards this goal in social psychology.

The CANAE System

The CANAE framework falls within a specific class of constraint satisfaction models: a content-addressable associative memory network. Notice that Dalege et al. (2016) assert that CANAE has two defining properties:

> ...correlations between evaluative reactions stem from pairwise interactions between the reactions [nodes] and, second, these interactions are aimed at optimization of the consistency of the evaluative reactions. (p. 5, paragraph 7).

Further, the attitude construct was introduced as the end-state of a dynamic memory system,

> ...i.e., the whole network of the evaluative reactions and the interactions between these reactions represents the attitude construct. (Dalege et al., 2016, p. 5, paragraph 5)

in which activation spreads across nodes:

> ..., information can flow from one variable to all other variables in a network. (Dalege et al., 2016, p. 5, paragraph 4).

Finally, observe that Dalege et al. (2016) invoke the Ising model from statistical physics to summarize their construction:

> In the CAN model, attitudes are conceptualized as networks of interacting evaluative reactions (e.g., feelings, beliefs, and behaviors toward an attitude object) and the dynamics of the networks conform to the Ising model. (Dalege et al., 2016, p. 6, paragraph 3)

The Ising model (see Brush, 1967; Cipra, 1987) inspired directly the early work on content-addressable associative memory models (for an historical view, see Anderson & Rosenfeld, 1989; Hinton & Anderson, 1989). The seminal work of John Hopfield in 1982 popularized the fact that content-addressable associative memory networks can be constructed in a manner that is equivalent to the Ising model. The reach of his work extended to the fields of computational neuroscience, cognitive science, artificial intelligence and physics (for an accessible exposition on this topic, see Chapter 2 of Hertz, Krogh, & Palmer, 1991).

Content-addressable associative memory networks[3] have well-understood properties that can be leveraged to understand their representational competencies and dynamics. First, they have a capacity, α, for storing patterns of inputs. Beyond this capacity, patterns start to interfere with one another. The attractors for patterns are stable end-states (fixed-points, ground-states) that summarize where such systems end up given an initial state vector (i.e., input to the system). These can be characterized as having a depth and an ability to deal with noise and damage (Hertz et al., 1991). They come in a variety of flavors to accommodate different assumptions about nodes (e.g., leaky, integrate-and-fire), activation functions (stochastic, deterministic) and

———

[3] We will use the terms attractor networks and attractor neural networks interchangeably with content-addressable associative memory networks.

other properties of the system (e.g., pseudo-inverse methods for increasing capacity). [4] In short, attractor neural networks have a rich history in computational neuroscience, cognitive science, artificial intelligence and physics of different varieties and come with clear, understandable results.

The technical details of CANAE implementation are as follows:

- There is a graph $G = G(V, E)$ consisting of a collection of beliefs (vertices from a set V) and relations between them (weighted edges from a set E).

- The state of vertex $i \in V$ is $x_i \in K_i$ where K_i is the state set for that vertex.

- For all i we have $K_i \in \{-1, 1\}$.

- The system state is $x = (x_1, x_2, \ldots, x_n)$.

- The system global energy H is defined using all $i \in V$ by
  $H(x) = -[\sum_{i \in G} \sum_{j \in N_G(i)} \tau_i x_i + w_{ij} x_i x_j]$ where $N_G(i) \subset V$ is the set of neighbors of i in G, *not* including i, w_{ij} is the weight of the edge $\{j, i\}$ and τ_i is the baseline parameter for vertex i (we use θ_i interchangeably with τ_i throughout the text). Assume that $w_{ij} = w_{ji}$.

- For $i \in V$ let $\sigma_i \colon \prod_{i=1}^n K_i \longrightarrow \mathbb{R}$ be the function defined by $\sigma_i(x) = H(x) - H(\bar{x})$ where x and $\bar{x}$ are configurations given the current and opposite state of vertex i, respectively.

- For each vertex i we define its vertex function as $\phi_i(x) = 1/(1 + e^{-\sigma_i(x)/t})$ where t is the temperature[5], a parameter of the system; this defines the probability that at any point in time a vertex i will flip to its opposite state: $P(c \longrightarrow o) = \phi_i(x)$.

A typical instance of CANAE is a discrete-time, asynchronous simulation. For each time step: (i) select a vertex i, (ii) compute $P(c \longrightarrow o) = \phi_i(x)$ and (iii) use $P(c \longrightarrow o)$ directly to decide if vertex i will change its state. Another common implementation is to draw n samples of the system state x from the Gibbs probability distribution (the Gibbs distribution is defined by the energe H over the configuration space (all possible states)). In our simulations below, we use a local computation for σ similar to Hopfield's formulation (Hopfield, 1982) which replaces the following functions in the CANEA implementation as stated above:

- For all i we have $K_i \in \{0, 1\}$.

- Define $\sigma_i(x) = \sum_{j \in N_G(i)} \tau_i + w_{ij} x_j$.

- The vertex function $\phi_i(x)$ defines the probability that the state of vertex i (x_i) will be equal to 1; in notation this is $P(x_i = 1) = \phi_i(x)$.

──────────

[4] See Gerstner, Kistler, Naud, and Paninski (2014) Chapter 17 and Trappenberg (2010) Chapter 8 for an expansion of these topics.

[5] We use $t = 0.001$ for Study 1 and $t = 0.1$ for Study 2 in this article.

Summary

We set out to verify two theoretical claims of CANAE: (i) local structural characteristics of individual components in an attitude network drive the global dynamics of attitude formation, and (ii) the network structure of attitude networks is driven by the trade-off between attitudinal accuracy and attitudinal consistency; the small-world structure serves this purpose. It is unknown whether or not these claims are logically consistent or derivable given the formal specification of CANAE as an Ising-like model. We will demonstrate that these two claims are, fundamentally, about the relation between attractors and network structure. The former issue is a question of the relation between attractor states and the relative position of nodes in the network. The latter issue is more subtle: it refers to the relation between network topology and its capacity in terms of the number of attractors supported by the system.

In *Study 1* we will demonstrate that specific predictions made by CANAE, with a particular data set (featured in CANAE), are too simplistic and do not characterize well the dynamics of the system. In *Study 2* we will use attractor network capacity to test the claim that a small-world network structure provides the right trade-off between the system's consistency and its accuracy. The small-world topology as specified by CANAE, provides a high degree of consistency but at the expense of capacity of the system (i.e., the number of attractors it will support). Discussion will follow.

Study 1

This study examined a set of theoretic predictions central to CANAE, predictions that typify how the psychological network literature interprets the relation between static psychological network properties and the dynamics of the associated systems[6]. To begin, we will look at the debut CANAE article (Dalege et al., 2016) which provided an estimate of an attitude network in reference to the U.S. 1984 presidential candidate Ronald Reagan using the American National Election Study (ANES) of 1984 (see Figure 2, right-panel in Dalege et al., 2016). Their predictions, using these data, emphasized node attributes:

> How a given node is connected in the network will influence whether and how change in this node will spread to other nodes. (Dalege et al., 2016, p. 10, paragraph 7)

More specifically, two network attributes of a node were singled out as important: membership in a cluster and node centrality. Two sets of nodes in the ANES-Reagan data were accompanied by exemplary node-level predictions: the nodes representing Reagan as setting a good example and whether he cares about his constituents, each of which is shown in respective order in the following two quotes:

> Thus, whether change in the negative affect cluster would spread through the network would depend on whether you change your mind that Ronald Reagan sets a good example. (Dalege et al., 2016, p. 11, paragraph 3)

---

[6] The network approach or network analysis approach in psychology is somewhat ill-defined (see Bringmann & Eronen, 2018); it usually, however, implies a certain perspective (e.g., vertices/nodes on a graph/network can affect other vertices via edges/weights) and set of methods (e.g., computation of network measures reflecting topology) that, taken as a whole are less ambiguous.

and,

> For example, the evaluative reaction with both the highest degree and
> highest closeness in the network of the attitude toward Ronald Reagan is the
> judgment of whether he cares about people like oneself. It is thus likely that
> ... change in this judgement would affect the attitude network to a large
> extent. (Dalege et al., 2016, p. 11, paragraph 5)

These predictions refer to the consequence or extent of effect of the perturbation
of single, individual nodes. The CANAE prediction, in its basic form, is that a node's
centrality and its cluster membership will relate to its extent of effect. Such a result
would lend support to viability of using static network structure properties to infer
something about the dynamics of a content-addressable associative memory system.
This notion is part of the CANAE canon (Chambon et al., 2022; Dalege, Borsboom, van
Harreveld, & van der Maas, 2017; Dalege, Borsboom, van Harreveld, Waldorp, &
van der Maas, 2017; Zwicker et al., 2020).

We posit the following criteria for testing this yet untested, general hypothesis: (i)
the test must invoke key aspects of the Ising model, either via numerical simulation or
mathematical analysis, (ii) the definition of extent of effect should reflect the notion of
perturbation, and (iv) the definition of extent of effect should be in relation to the likely
attractors or fixed-points of the system. These criteria offer a fair minimum set for
testing this hypothesis.

To date, the work on CANAE that follows most closely to these criteria is in
Dalege, Borsboom, van Harreveld, and van der Maas (2017). The principal result was a
difference in the sum score, defined as $\Sigma_i x_i$ (see Introduction for definition of x_i; this is
a system-level state of all vertices), between two conditions–one condition perturbed the
most central vertex by forcing its τ_i to a value of 1; the other condition did the same for
the least central vertex. The difference in sum score was approximately 2 points; the
mean of 1.18 (SD= 7.99) for most central vertex and a mean value of -1.04 (SD = 7.84)
for the least central vertex. The operational range in sum score in this demonstration
was on the order of 21 points (integer values from -10 to 10).

The use of sum scores as the point of comparison doesn't reveal much about the
extent of effect a vertex has on the system when perturbed. Of far greater value would
be understanding the extent of effect in respect to the attractors of the system,
something that was knowable given the small size of the system studied by (Dalege,
Borsboom, van Harreveld, & van der Maas, 2017)[7]. In short, we don't have a
meaningful way to interpret the difference in sum scores without more knowledge of the
system's characteristics in terms of stable attractor states.

The simulation code and the links to the original empirical data are available in
the supplemental information associated with Dalege, Borsboom, van Harreveld, and
van der Maas (2017). These data were a subset of the 2012 American National Election
Survey consisting of 10 questions about the presidential candidate Barack Obama. Two
vertices (items) were identified to be the most influential in terms of their extent of
effect. These were in reference to Barack Obama's honesty and leadership qualities:

> That the node Led[ership] is closely connected to all communities also
> explains why it is the node with the highest closeness. Change in this node is

---

[7] The system under study in (Dalege, Borsboom, van Harreveld, & van der Maas, 2017) was composed
of 10 vertices; thus, the number of unique system states was 1024

thus likely to affect large parts of the network. Hns[honesty] has the highest strength because it has strong connections to ...[other nodes]. Change in the node Hns[honesty] would thus strongly affect many other nodes. (Dalege, Borsboom, van Harreveld, & van der Maas, 2017, p. 532, paragraph 3)

Using the code provided by the authors (Dalege, Borsboom, van Harreveld, & van der Maas, 2017), we expanded their study by perturbing each of the remaining eight vertices (they perturbed two of 10) as a kind of prelude to Study 1. We added a more informative measure (Wasserstein distance) and showed the distributions of sum scores (in addition to the mean) across all 10 vertices in comparison to the baseline condition used in their study.

Figure 1 shows the distributions of sum scores for each perturbation condition. We see two gross features: (i) the baseline condition in Dalege, Borsboom, van Harreveld, and van der Maas (2017) reveals one very frequent attractor (when all vertex states x_i's were -1), and (ii) the 10 perturbation conditions are very similar in their distributions of sum scores. Included in each panel of Figure 1 are the Wasserstein distance (of the perturbation distribution compared to the baseline distribution) and the mean of the sum scores.

The original article also provided code for computing the centrality of each vertex. Using this information Figure 2 illustrates the relation between vertex centrality and extent of effect (provisionally defined by the mean sum score or by the Wasserstein distance between the perturbation distribution of sum scores and the baseline distribution). The manifest feature of this figure is that there is very little variation in extent of effect across vertices, a pattern that matches the distributions shown in Figure 1; further, no clear relation is exhibited between vertex centrality and extent of effect.

Our detailed analysis of Dalege, Borsboom, van Harreveld, and van der Maas (2017) serves three purposes. First, it provides a synopsis of the most relevant prior CANAE work on the problem of understanding the effects of perturbation of vertices and the extent of effect of such perturbation on attitude networks. Second, it highlights the general method we employed in Study 1: define extent of effect of perturbations of vertices (we do it in a rigorous manner below) and look at the relation between extend of effect and the centrality of vertices. Third, it offers a preliminary result: the extent of effect may not be related to the centrality of vertices.

Other work in the CANAE canon, both empirical and simulation based, did not meet our criteria for testing the hypothesis in question. Of the studies that directly addressed the extent of effect of perturbation, some were correlational and thus did not offer the control required for a perturbation study (Chambon et al., 2022; Dalege, Borsboom, van Harreveld, Waldorp, & van der Maas, 2017; Zwicker et al., 2020). Of these, one included simulation but did not sufficiently represent our criteria because it lacked perturbation methods (Dalege, Borsboom, van Harreveld, Waldorp, & van der Maas, 2017). Some work in the CANAE canon addressed other important issues, but did not broach the topic of vertex perturbation and extent of effect (e.g., Dalege et al., 2018, 2019; Dalege & van der Maas, 2020).

Now we move to the principal work of Study 1. The objective of this study was to demonstrate, with a detailed example, a rigorous method for testing the CANAE hypothesis of the relation between vertex centrality and extent of effect using the original predictions and original data from Dalege et al. (2016). The use of a realistic attitude network, derived from empirical data using the method prescribed by CANAE makes up for a lack of generality (something we address in the General Discussion) with

a degree of relevance and complexity. Keep in mind that Study 1 is telling a story from a related but different technical perspective than that used in the CANAE literature; it is focused on stable system states (attractors) and how perturbation might affect said states; it is also quite detailed. We are convinced that the hypothesis and its demonstration require such detail for full understanding. The language, thus, may seem foreign to readers versed in CANAE.

The method was simple in principle.[8] First, we derived a referent set of attractors for the system–those that reflect the distribution of natural states of the system. Second, we perturbed the system to test the degree to which the perturbation generated change in comparison to the distribution of the referent attractors. The method of perturbation, described in detail next (simulation *Sets 1* and *2*), fixed the states of individual nodes, one at a time, throughout the full time-course of a simulation. Third, for each node, we computed a summary measure of its extent of effect as the difference between the referent and perturbation distributions. From this summary measure, we could see the relation between the node-level network properties (centrality and cluster membership) and the node-level summary measure of extent of effect.

Prior to our simulation results, we can provide a preliminary and coarse-grained prediction based on the analytic results of a simple case: the stability analysis of an asynchronous Hopfield model with only one attractor state ξ_i (one stored pattern) in which the thresholds $\theta_i = 0$, the states $\in \{-1, 1\}$, $w_{i,j}$ restricted to $\in \{-1, 1\}$ and the sign function is used for the node state dynamics[9]. Borrowing from Hertz et al. (1991), we know that if the initial state S_i is equivalent to the attractor state, $S_i = \xi_i$, then the system will likely not change. Further, if more than $1/2$ of the nodes are correct, meaning more than $1/2$ of $S_i = \xi_i$, then the system will likely settle to ξ_i as the system will correct the incorrect bits in S_i. In reference to the Dalege et al. case, flipping one bit from S_i would have near zero effect on the other nodes if the system was in or near its attractor state ξ_i. Finally, notice that flipping only one bit (from state ξ_i) would nearly guarantee, if you let the system evolve, that the bit would flip back to its original state the next time it was updated with probability 1 and, since no other bits would change, the system would settle to ξ_i. It is also true that if nearly $1/2$ (or greater) of the bits were different from ξ_i, then the flipping of one bit k would likely have an effect. Notice that the latter case is the only case that predicts perturbing one node would make a difference in the system state.

What can we learn from this simple analytic case? The effect of flipping a bit depends on the state of the system. In this simple case, when the system is in or near its attractor, flipping a bit will have a near zero extent of effect. In fact, the system must be far from its attractor for the flip of one bit to have an effect.

Simulations (*Set 1* and *Set 2*) used the dynamic network as described in the Introduction (we will use the term Hopfield network for our model henceforth). For both sets of simulations the weights w_{ij} and baselines τ_i of our Hopfield network were fixed to those provided in the ANES-Reagan network used in the original article (Dalege et al., 2016). We manipulated the initial conditions (i.e., the initial state vector), any perturbation constraints (e.g., fixing a node's state to specific values) and the temperature. We will borrow the term "evaluative reactions" from the original article to mean nodes in the network.

––––––––

[8] We will use the terms node and vertex interchangeably.

[9] For this simple case we borrow notation from (Hertz et al., 1991)

In both sets of simulations, each of the 22 evaluate reaction nodes was perturbed separately, generating a node-specific subset of simulations in which the state of the node was fixed to "on" (state $= 1$) for the entire simulation period. *Set 1* differed from *Set 2* only in terms of how we constructed the initial conditions. For *Set 1*, each node was tested across all possible combinations of states as initial conditions (2^{22} initial conditions). Thus, each node was tested in 4,194,304 simulations; this resulted in 92,274,688 ($22 \times 4,194,304$) simulations in total for *Set 1*. In *Set 2* we tested the effects of a single node under the conditions of defining the remaining elements of the initial condition state vector as all zero. For *Set 2* we ran 1,000 replicates per each node; this resulted in 22,000 (22×1000) simulations in total for *Set 2*. Although *Set 1* and *Set 2* were similar in terms of methodology, the former focused on the degree to which sustained input from the node affected the system across the full set of possible initial conditions while the latter emphasized the degree to which sustained input from a single node could push the system's dynamics from the zero attractor state.

We recovered the system's referent attractors using our Hopfield network. We fixed the temperature to a very low value and ran all possible combinations of states as initial conditions (2^{22} initial conditions)[10]. In short, we ran 4,194,304 simulations, each representing a unique state vector as the initial condition and for 2000 discrete simulation time-steps. The results revealed 15 attractors, shown in Figure 3. The relative frequency of each attractor, in reference to its Hamming distance from the most frequent attractor, is provided in Figure 4; Figure 5 provides the same for energy. From these results we can summarize the system under a low temperature condition as having a single low-energy and highly frequent attractor with another set of three somewhat frequent attractors, one of which is close in energy and in distance to the primary attractor. The other two attractors are more distant and higher in energy from the most frequent attractor and close to one another in both energy and Hamming distance. The remaining 11 referent attractors are infrequent.

Figure 3 reveals some interesting features of this system (labels of attractors are provided in decimal values for the binary values of the state vectors): (i) the most frequent (4194218) and the second most frequent (85) attractors were full complements, (ii) attractor 85 was composed of only and all the negative valence nodes in the system, (iii) the third and fourth most frequent attractors (139349 and 4191598, respectively) were mixtures of the first two most frequent. The attractors of the system are well-structured and should serve well as a baseline for the perturbation analysis.

Our main results relied on, for each node separately, a comparison between the distribution of attractors under perturbation to the distribution of referent attractors, what we dubbed *consequence* or *extent of effect* of the perturbation. Formally, we defined the node k extent of effect, E_k, in the following way. First we define $dist_r$ (r for referent) as the discrete categorical frequency distribution of the 15 referent attractors. Let us treat $dist_r$ as a vector and $r = dist_r/\|dist_r\|$ to be its Euclidean normed vector. Then, the node k extent of the effect $E_k = \|r - s\|$ where s is the normed vector of a comparable frequency distribution[11] generated by the set of perturbation simulations for the node in question.

The measure E_k captures the Euclidean distance between the referent distribution and the node specific distribution, the latter generated via perturbation. However, it is

———

[10] We compared the simulations presented here with a higher temperature and found similar results.

[11] By comparable we mean the representation of the frequencies of the same 15 referent attractors.

not guaranteed that, for any node, any of its fixed-points under perturbation will match exactly (e.g., a Hamming Distance of 0) any of the referent fixed-points. So, we devised a procedure for computing $dist_s$ and s the description of which is provide in Appendix A. In the process, we introduced a measure H_k that served as an essential quality check on $dist_s$ using Hamming Distance to gauge how similar the node specific attractors were to the referent attractors.

In summary, for each node k we computed two measures: E_k was the extent of effect on the network; H_k was a quality check. Our primary result will show the relation between node centrality, clustering and E_k. We use H_k as a sanity check–if it is to high, any results found given E_k will be less convincing because the node specific attractors would not be similar to the referent attractors.

Before we look at the primary result it will be informative to compare the referent attractor distribution ($dist_r$) and the perturbation distributions for each node k ($dist_s$). Figures 6 and 7 show this comparison for simulations *Set 1* and *Set 2* respectively. In general, for simulation *Set 1*, there was reasonable alignment for all nodes in respect to the referent fixed-point distribution, with some clear and systematic differences: (i) perturbation in the negative cluster nodes generally aligned better than the most central node, forced a reduction in the frequency of the most frequent referent fixed-point (4194218), and boosted slightly the frequency of the second most frequent referent fixed-point (85), (ii) perturbation in the most central node forced a reduction in referent fixed-point 85 and boosted fixed-point 4194218, (iii) perturbation of some nodes, aside from both the most central and those from the negative cluster, resulted in very high frequency for the most frequent attractor. For simulation *Set 2* the patterning was different: (i) perturbation of the most central node was the only node that shows substantial frequency for the most frequent referent attractor (4194218), but most of its frequency matched attractor 85, an attractor that does not contain the perturbed node itself; (ii) perturbation of the negative cluster nodes drive the system to referent attractor 85; (iii) perturbation of some of the other nodes result in high frequency of the referent attractor 85, even though these nodes are not part of that attractor.

In this preliminary comparison, simulation *Set 1* suggests that perturbing the network across all possible initial states does not have much of a effect on the distribution of attractors while simulation *Set 2* shows that many of the perturbations settle into fixed-point 85. The latter result indicates that under low energy initial conditions (just one initial node state = 1), the fixed-point 85 is the main attractor, probably attributed to the fact that three of the four nodes that define this attractor had the largest baseline values (a stronger push to be "on"; ≥ 1.4 standard deviations above the average)[12].

We now move to the primary result of *Study 1*–to understand the extent of effect on the system of central and negative clustered nodes as hypothesized in the original CAN article. The summary measure E_k provides a direct understanding of how perturbations of nodes along the spectrum of centrality and clustering affect the system in comparison to the referent fixed-points. Figures 8 and 9 show the relation between node centrality and E_k for simulations *Set 1* and *Set 2* respectively and provide tags for the negative cluster nodes. The picture is clear for both sets of simulations: There is little relation between centrality or negative cluster membership and E_k. *Set 1* differs from *Set 2* in two ways. First, the perturbations in simulation *Set 2* had a much larger

---

[12] Baseline values of nodes capture the probability of the node state = 1 sans input from other nodes.

effect on E_k on the whole compared to *Set 1*. Second, the most central node was more unique (compared to the other nodes) in its value of E_k for simulation *Set 2* but not for *Set 1*; it was closer to the referent distribution, something that can be seen in the distribution shown in Figure 7 for the most central node.

Figures 10 and 11 show the relation between node centrality, negative cluster membership and H_k for simulations *Set 1* and *Set 2* respectively. We show this to provide evidence that the measure E_k was based on reasonably small Hamming distances. (For *Set 1* the Hamming statistics were: mean=0.06, std=0.07, min=0, max=0.34; for *Set 2*: mean=0.64, std=0.36, min=0, max=1.49).

In summary, the primary results using E_k as a summary measure of the extent of effect of a node was that neither node centrality nor membership in the negative cluster were related to E_k. Our claim should be considered in the context of the patterns of the distributions $dist_s$ in *Set 1* and *Set 2* where we did see some systematic effects, albeit small, when perturbing the most central node or those with membership in the negative valence cluster.

Transparency and Openness

In reference to the empirical data used in Study 1, as per our objective of testing the CANAE model predictions as given in (Dalege et al., 2016), the sample size, any data exclusions and construction of the attitude network were identical to those used in the study and were, thus, not in our control. All simulation and analysis code (and instructions for generation of the original data as used in the target article (Dalege et al., 2016)) are provided at https://github.com/mark-orr/Causal_Attitude_Network_Model_Comment. All analysis and simulations were conducted within a conda virtural environment. The original data is easily obtainable by anyone once completing a simple registration process on the data holders web portal. The specification file for this conda environment is included in this Github repository for reproduction of the environment. This study's design and its analysis were not pre-registered.

Discussion: Study 1

Study 1 did not support the general CANAE assertion that a node's extent of effect when perturbed was directly related to its centrality or cluster membership. Neither node centrality nor cluster membership was related to nodes' extent of effect on the system. We did see minor, systematic and reasonable system-level effects on the distribution of the perturbed states, especially in simulation Set 2, but these were driven by the high probability of ending up in attractor 85 when starting very near the zero attractor, not by an extensive effect on the system. It seems that the viability of using static network structure to infer system dynamics, at least in this specific case, is low.

Study 1, as a case study, doesn't readily offer generality. We claim, however, that it does offer a general lesson, one that could and should be minded when studying psychological networks. Predictions of system-level dynamics should respect the nature of the system, and when possible, leverage existing results and methods for understanding the system-level dynamics. Recent work on the issues surrounding interventions based on static network properties in clinical psychology (e.g. Bringmann, 2021; Bringmann et al., 2019; Haslbeck et al., 2021) makes a related claim. The overarching issue in the use of psychological networks in clinical psychology lies in a

lack of scientific understanding–we don't yet know enough about the inter-nodal causal processes in clinical psychology to construct formal models of the relevant system processes and thus cannot properly study their dynamics (Bringmann, 2021; Bringmann et al., 2019).

Notice that this limitation does not apply to attitude networks–in both the CANAE model and other prior work with neural networks (Conrey & Smith, 2007; Monroe & Read, 2008; M. G. Orr et al., 2013) the system is specified completely and to a degree that readily affords simulation or analytic results. Yet, the predictions put forth in the CANAE canon (Chambon et al., 2022; Dalege et al., 2016; Dalege, Borsboom, van Harreveld, & van der Maas, 2017; Dalege, Borsboom, van Harreveld, Waldorp, & van der Maas, 2017; Zwicker et al., 2020) were largely divorced from the theory of Ising-like systems, theory that provides not just results and methods but, more importantly, how to think about the class of problems. We offer a quote from a classic work in Ising-like systems (Hertz et al., 1991):

> ... we are usually not satisfied with simply stating or deducing a given result, but instead try to show the reader how to think about it, how to handle and hold it. (Hertz et al., 1991, p. XX, paragraph 2)

Our results were very basic, yet illustrative of the possible analytic and synthetic approaches available in the dynamical systems literature used to address questions of network structure and system dynamics (see H. Mortveit & Reidys, 2007, for a detailed introduction). To emphasize this point, we offer an extension of the stability analysis we provided in the beginning of Study 1 (based on (Hertz et al., 1991)) which predicted that perturbing a single node will likely not have a strong extent of effect on the network dynamics, a prediction generated from a highly simplified case. The simulations in Study 1 were more complex in terms of the process dynamics defined at the node-level and in the set of weights w_{ij} and baseline parameters τ_i (in place of thresholds θ_i) which reflected real-world, correlated data. This notwithstanding, we can use a similar analysis to understand the relation between fixing a node and its extent of effect on the system. Notice that fixing the state of node k, x_k, for all time-steps of a simulation, as we did in simulations *Set 1* and *Set 2*, is nearly equivalent to a simulation in which all nodes but k are present and a constant C_i (to stand in for node k) captures the effect of node k on the state x_i of each node i (C_i is the vector of constants $\{i = 1, ..., (n-1)\}$ where n is number of nodes in the network and each value is equal to $w_{ik}x_k$). Then, using this system, the question of the extent of effect of a node is reformulated[13]: Are the attractors of the system different with or without C_i? We can address this question using an approach similar to the highly simplified case used at the beginning of Study 1.

From the Introduction, we can see that for any node i, the input $\sigma_i(x)$ would now be $\sigma_i(x) = \sum_{j \in N_G(i)} \tau_i + w_{ij}x_j + C_i$. We can then generate from the original CANAE article ANES data (Dalege et al., 2016) the parameters to make a first approximation of the process dynamics of the average node using the expected values were $E[w_{ij}] = 0.19$, $E[x_j] = 0.61$, $E[\tau_i] = -1.92$ and $E[C_i] = E[w_{ij}]$. This formulation allows for a simple comparison of $\sigma_i(x)$ with and without C_i which were 5.86 and 6.47, respectively. Both of these values when plugged into the node update rule $\phi_i(x)$ are approximately one. That is, under average conditions, with and without C_i the probability P_i that any node

---

[13] This reformulation is not fully equivalent to simply fixing k versus not fixing k because the latter does not necessarily mean that $k = 0$; we use this example as an illustrative simplification.

state x_i will be "on" (in state $\{1\}$) is one. Thus, flipping a single node "on" or "off" will have little chance of having an effect on the network, ceteris paribus, when conditions are typical. Notice that this preliminary result accords with the frequencies of the most frequent referent attractor in the system (see Figure 6) where most $x_i = 1$.

The point of this analytic exercise is to bring clarity to the problem of whether fixing a node in this system will likely have an extensive effect. Stability analysis, in its analytic form, is one fruitful way to think of the problem because it captures the node-level dynamics, but leaves out the details of the network structure [14]. Simulations *Set 1* and *Set 2* capture both the node-dynamics and the network structure. Taken together, these results emphasize that such questions benefit when accounting for all of the key properties of the system, something that the work in the CANAE canon that relates to extent of effect (Chambon et al., 2022; Dalege et al., 2016; Dalege, Borsboom, van Harreveld, & van der Maas, 2017; Dalege, Borsboom, van Harreveld, Waldorp, & van der Maas, 2017; Zwicker et al., 2020) does not incorporate. CANAE asserts that attitudes are an Ising-like systems but does not take into consideration the formal meaning of this class of system in its analysis of extent of effect.

Study 2

The CANAE model makes a unique theoretical assertion: attitude networks conform to a small-world network topology. In the authors' words:

> In the CAN model, the structure of attitude networks is held to conform to a small-world structure (e.g., Watts and Strogatz, 1998): Evaluative reactions that are similar to each other form tight clusters, which are connected by a sparser set of "shortcuts" between them. (Dalege et al., 2016, p. 3 , paragraph 4)

This structure results from a need to satisfy a trade-off between consistency and accuracy. In the authors' words:

> Attitude networks are driven by the trade-off between optimization (i.e., consistency between evaluative reactions) and accuracy. This trade-off results in a small-world structure, in which evaluative reactions, that are similar to each other, tend to cluster. (Dalege et al., 2016, p. 14, paragraph 7)

Consistency is driven by a patterning of excitatory and inhibitory weights in the network:

> To acquire a consistent state, evaluative reactions of the same valence generally have excitatory influence between them and evaluative reactions of different valence generally have inhibitory influence between them. (Dalege et al., 2016, p. 6, paragraph 3)

And, accuracy is afforded by both clustering and the sparsity of small-world connections between clusters, the combination of which are key structures inherent in the small-world topology:

————

[14] This analysis was at the level of individual nodes; it doesn't directly tell us how this might cascade without further and detailed analysis (see Chapter 2, section 2.5 of Hertz et al. (1991)).

> Clustering allows for energy reduction within clusters (e.g., all evaluative
> reactions toward a person that pertain to the dimension of warmth are
> highly aligned) but also allows for accuracy by having unaligned or even
> misaligned clusters that do not cost much energy (e.g., the evaluative
> reactions that pertain to the dimension of warmth are not highly aligned to
> the evaluative reactions that pertain to the dimension of competence).
> (Dalege et al., 2016, p. 6, paragraph 4)

In summary, the small-world structure is hypothesized to support the drive for the
trade-off between consistency and accuracy by virtue of a particular configuration of
weights: a dense network of excitatory/positive weights among similar evaluative
reactions and a sparse network of inhibitory/negative weights between non-similar
evaluative reactions. The small-world topology, combined with the prescribed pattern of
excitatory and inhibitory edges, is hypothesized to be the sweet-spot in terms of the
supposed trade-off. This intriguing theoretical hypothesis had been neither empirically
tested nor formally verified, yet it persists in the CANAE canon (e.g., see Zwicker et
al., 2020).

It is difficult to evaluate whether the small-world supports the hypothesized
consistency/accuracy trade-off because, although consistency was defined
unambiguously (as energy), accuracy was not. This study attempts to bring clarity to
this issue. We start with the notion of accuracy.

In the CANAE model, accuracy was invoked to account for the possibility of
non-aligned clusters emerging through the sustained process of learning about and
experience with an attitude object. This would, presumably, pressure the network to
represent dissimilar clusters of evaluative reactions that reflect different
dimensions/aspects of an attitude object (most likely valence, but other dimensions
might be candidates). Thus, the CAN notion of accuracy refers to the capability of the
system to afford the representation of multiple non-similar or unaligned clusters.
Simply put, it reflects the world as it is, typically.

This description of accuracy, however, is incomplete. It does not distinguish
between a system with one dominant attractor that represents unaligned clusters
simultaneously and a system with a set of attractors each member of which represents
an aligned cluster. By definition, the latter case is non-simultaneous because an
attractor is a fixed-point of the system. Introducing the notion of simultaneity implies
two categorically different kinds of accuracy, an important distinction because each kind
imposes its own requirements on the system. Simultaneous accuracy requires that the
system has a stable state in which all evaluative reactions are endorsed.
Non-simultaneous accuracy, assuming there exist unaligned clusters, requires at least
two attractors that discriminate between them to some degree. This issue has not been
addressed in the CANAE canon (e.g., Dalege et al., 2016; Zwicker et al., 2020), its
importance, as we will demonstrate, notwithstanding.

The distinction between simultaneous and non-simultaneous accuracy is
unnecessarily burdensome in its complexity. A simpler notion that serves equally well in
this context is capacity. In the Hopfield model, capacity reflects the number of
attractors stored in a system[15]. So, instead of evaluating whether the small-world
network structure satisfies a consistency/accuracy trade-off, we can ask: What is the

---

[15] See Hertz et al. (1991) for precise, formal treatment of capacity in Hopfield models. Our use of the
term is less formal; we define it to only mean the number of attractors supported by the system.

capacity of a small-world CAN model? Does the capacity differ in comparison to more regular or random networks? And, how do capacity and consistency relate?

We begin addressing these questions by way of a simple demonstration, one that was inspired by a canonical neural network model: the Necker cube visual illusion (Rumelhart, Smolensky, McClelland, & Hinton, 1986). The Necker cube has two kinds of attractors. One kind attractor represents activation of either one or another cluster of nodes as "on" which we dub, borrowing from the Necker cube model, left- or right-facing. The other kind attractor is defined when both clusters are "on" simultaneously; in terms of the Necker cube, both clusters "on" is an "impossible" representation of the cube. The mapping to accuracy, as discussed above, is straightforward if we assume that the clusters are misaligned: a system that exhibits attractors for the left- or right-facing state shows non-simultaneous accuracy (capacity > 1); a system that settles to the the "impossible" interpretation, in which both clusters are active simultaneously, reflects simultaneous accuracy (capacity of one). Next, we demonstrate that the network structure dictates the capacity it affords.

We simulated a pseudo-Necker cube using the same Hopfield model as used in *Study 1* but with 8 nodes, τ_i values of zero, and symmetric weights $\in \{-1, 1\}$. Figure 12 shows the basis for four different Hopfield simulation sets (*Set* A - D) each defined by one of four configurations of 12 weights (labeled A - D in Figure 12). Across simulation sets, we systematically increased the proportion of negative edges, which only connected the two clusters (Panel A shows the two clusters clearly.) We ran 100 simulations for each combination of the weight configuration and three initial state vectors (for the latter we used *random* with $p(x_i = 1) = 0.50$, all ones, and all zeros) and computed the measures shown in Table 1.

The manipulation of the dependent variables (proportion impossible and mean energy) is represented by the proportion of negative edges in Table 1 and by the graphical depiction in Figure 10. Set A had zero negative edges and thus, no connections between the two clusters. Moving from Sets B to D, we increased the number of negative edges by two. These negative edges were constructed so that all nodes always had the same degree (number of connections).

The first result was that the proportion of runs that yeilded the impossible attractor dropped dramatically from one for Set B to close to zero for Set C and D. In sum, Sets A and B settled on the impossible attractor and Sets B and C largely settled on either the left- or right-facing attractor. The second result was that mean energy increased nearly linearly with the proportion of negative edges in the network.

In this demonstration, we show that the essential structure generated by the hypothetical CAN learning mechanism–dense clusters of nodes connected by positive/excitatory weights in conjunction with sparse, between-cluster negative weights–resulted in one attractor state, albeit it at a low energy. When the clustering is broken, the system yielded two complimentary attractors (likely driven by between-cluster inhibition), but with comparatively higher energy.

Given this demonstration we can revisit the trade-off between consistency and capacity and guess where the CANAE model sits in respect to it. Set B seems, to a first approximation, most like what the CANAE model proposes. It has sparse inhibitory connections, is relatively low in energy and high in consistency, and represents both clusters simultaneously. Further, we suggest that Sets C and D do not satisfy the CAN trade-off because the energies are relatively high.

In our demonstration, consistency was inversely related to capacity. To increase

consistency, you must trade-off the systems capacity. The hypothesized trade-off (per Dalege et al., 2016), if we must have one, is not between consistency and accuracy but between consistency and capacity. The CANAE model trade-off is where we have high-consistency and low-capacity. In reference to the CANAE notion of accuracy (Dalege et al., 2016), it seems that it is of the simultaneous kind.

Simulations *Set 3* and *Set 4* put these intuitions to the test in the context of the formal specification of the small-world graph generating algorithm as defined by Watts and Strogatz (1998); these are a direct extension of the pseudo-Necker cube demonstration but at a much larger scale, one that captures better the formal graph properties of the small-world phenomena.

Each simulation set ran a set of Hopfield models (same general specifications as in *Study 1*) each with 1000 nodes, τ_i values of zero, states $\in \{0, 1\}$, symmetric weights $\in \{-1, 1\}$. The only difference between simulation *Set 3* and *Set 4* was the initial state vector; for *Set 3* it was ones and for *Set 4* it was random (defined in the same was as for the pseudo-Necker cube demonstration). The graph for each Hopfield model was generated using the Watts-Strogatz algorithm (Watts & Strogatz, 1998). The rewire parameter r (the probability of rewiring each edge) controls the small-world regime, defined roughly between $r = 10^{-2}$ to 10^{-1}. At the end-points of the rewire parameter (no rewiring and probability of rewiring $= 1$), the graph is considered regular and random, respectively, in its topology (Watts & Strogatz, 1998). In our simulations, we used an extension of the small-world rewiring algorithm such that all rewired edges are negative and all non-rewired edges are positive. In effect, this means initializing with a regular (ring-lattice) graph g with $k = x$ (x is the number of neighbors for each node) and all positively weighted edges. Then assign -1 to any edge that is rewired.

The point of this rewiring scheme was to capture the core notion of the CANAE model network topology. In the small-world generating algorithm, rewired edges are considered as bridges between clusters. By asserting that they are negative, we capture the CAN notion of inhibition between clusters. This was analogous to the method we used for the pseudo-Necker cube demonstration.

Both simulation sets spanned the ordered set $R = \{0, 0.001, 0.005, 0.01, 0.05, 0.10, 0.20, 0.40, 0.60, 1\}$ where each value reflected the rewire parameter in the Watts-Strogatz small-world model (Watts & Strogatz, 1998). For each simulation set, we ran 10 replicates of the Hopfield model for each value of R [16]. This generated 100 simulations per simulation set. The idea, just as in the pseudo-Necker cube demonstration, was to understand the relation between consistency (as energy) and capacity of the network within and outside of the small-world regime (approximately an r value between 10^{-2} and 10^{-1}). The predictions for our simulations come directly from the results of our pseudo-Necker cube demonstration: when the probability rewire is in the small-world regime, the energy of the system should be relatively low and only one attractor should be found.

Figure 13 shows the primary result for simulation *Set 3* in terms of energy. As predicted, the small-world regime (between rewire probability of 10^{-2} and 10^{-1}) has nearly the minimum energy compared to higher rewire probabilities. Figure 14 shows the primary result in terms of capacity. Here, we see the attractors for all of the *Set 3* simulations, annotated by the value of the rewire parameter r. The system is stable, in

––––––––

[16] A replicate was defined as one graph generation using the small-world algorithm; thus, the 10 replicates were likely not identical graphs.

respect to its initial state x until $r \geq 0.20$. After this point, the system exhibits multiple attractors across replicates.

Figures 15 and 16 show the results for simulation *Set 4*. The results show a close correspondence with those from simulation *Set 3* even given a random initial state vector.

For both simulation sets, in the small-world regime, the dynamic was driven by a preponderance of positive, excitatory weights within clusters and a small fraction of negative, inhibitory weights across clusters, a structure that yields little inhibition between clusters of evaluative reactions that are not similar. The effect was that all clusters eventually became active, due to the combination of mutual reinforcement within clusters (from the excitatory/positive weights) and the stochastic nature of the node dynamics (each node, even without input, will change states with some probability). This was the case for both initial conditions as was demonstrated by the similarity of results between simulation sets *Set 3* and *Set 4*.

In sum, the predictions of the pseudo-Necker cube demonstration extended to these two simulations. In the small world regime, the system had relatively low energy and a capacity of one.

Transparency and Openness

The data in this study were completely synthetic and theoretical. Thus issues of sample size, any data exclusions and construction of constructs were not in relation to human data but in accord with reasonable practices in simulation and analysis of theoretical dynamical systems. All simulation and analysis code (and instructions for generation of the necessary data are provided at https://github.com/mark-orr/Causal_Attitude_Network_Model_Comment. All analysis and simulations were conducted within a conda virtural environment. The specification file for this conda environment is included in this Github repository for reproduction of the environment. This study's design and its analysis were not pre-registered.

Discussion: Study 2

By the CANAE model, the drive or need of attitudinal systems to reach a trade-off between consistency and accuracy is axiomatic (Dalege et al., 2016). Our issue is not with this axiom but with the lack of clarity in the definition of accuracy. This made it difficult to evaluate the claim that the small-world topology, with the right patterning of excitatory and inhibitory weights, supports the trade-off.

Upon analysis of the issue, we found it more natural to state the trade-off in terms of capacity. Networks with excitatory, dense clusters that inhibit one another with sparse networks exhibited high consistency but have a capacity of only one; increases in capacity reduce consistency. This pattern was clearly demonstrated in the pseudo-Necker cube simulation and simulation *Sets 3* and *4*. If there is a trade-off in relation to the CANAE model, it is between consistency and capacity.

From multiple disciplinary perspectives (e.g,. neural networks, neuroscience, machine learning, and cognitive science) the behavior of the small-world CANAE is unusual. Its representational function is, by virtue of the small-world assertion, to store one attractor in which all features it has learned about an attitude object are activated in all contexts. It does this using a specific network structure–one that essentially

minimizes the inhibition in the network while maximizing the excitation–and a specific dynamic, that of an Ising-like system. The unusual part is that, because of its low capacity, it doesn't leverage the useful aspects of content-addressable memory networks (e.g., context sensitivity, graceful degradation). It is fair to ask: what is the purpose of a network attitude model with Ising-like dynamics and a capacity of one? We will address this issue further in the general discussion[17].

In summary, we see in this study a similar issue with CANAE as found in Study 1. A focus on network structure without thorough consideration of the interplay between network structure and the dynamics of the system.

General Discussion

We will focus the general discussion on two topics. The first is theoretical: What is the function of the CANAE system? There is enough ambiguity in the CANAE literature about this question to render its answer somewhat opaque. The second is practical: How should we make predictions in respect to the dynamics of psychological networks? We will explore a different but prominent computational model of attitudes, the Attitudes as Constraint Satisfaction (ACS) model (Monroe & Read, 2008), to address this latter question.

What does CANAE do?

The theoretical capacity of a network of 1000 vertices is about 130 states (Hertz et al., 1991). Within the CANAE small world regime, as shown in Study 2, the capacity of a network of the same size was one. This limited capacity was likely due to the peculiar weight distribution dictated by CANAE in which excitatory weights within clusters vastly outnumbered inhibitory weights between clusters. In the Discussion of Study 2, we raised the question: What is the purpose or function of such a limited capacity system for the agent or person in terms of evaluation of an attitude object? In attempting to answer this question, we took a detailed, in-depth look at the CANAE literature. Our findings revealed something obscure in the CANAE literature. CANAE evolved in terms of both the meaning of vertex states and the typical kind of distribution of the network weights. In fact, CANAE split into two separate forms–a content-addressible associative memory network and a cusp-catastrophe network–each with unique characteristics and markedly different functions. The surfacing of these forms provided a resonable answer to our question about CANAE's function. It also raised other critical issues, which we will explore after we answer our initial question.

Evolution of the Function of CANAE. CANAE can be conceptualized as a content-addressable associative memory system (we provide such a framing in the Introduction). In this form, CANAE asserts that the meaning of vertex states x_i is endorsement or not of the evaluative reactions they are meant to represent. The meaning of the weights w_{ij} is asserted as the causal influences among the set of vertices. Inhibitory as well as excitatory weights are expected in a network for which the evaluative reactions do not align (e.g., are of different attitudinal valences); ultimately, it is the social context from which one learns that drives the distribution of weights. The putative function of such a system is to store and retrieve system states that reflect

---

[17] If we would have used a state set $K_i \in \{-1, 1\}$ in simulation *Sets 3* and *4*, a reverse state would exist that is the logical complement of the one stored attractor. Thus, when we say one attractor we mean for it to include whatever reverse state exists.

exposures in the social environment or through self-reflective processes. This form captures how CANAE is described in its debut (Dalege et al., 2016).

The semantics of CANAE were altered two years later, or maybe more accurately, another CANAE form was defined. In Dalege et al. (2018) it was suggested that CANAE is akin to the cusp catastrope model of attitudes (Liu & Latané, 1998). In this form, the meaning of vertex states x_i was valenced:

> The CAN model can easily integrate the catastrophe model of attitudes...Thresholds in the CAN model directly relate to the valenced information a person receives regarding an attitude object... (Dalege et al., 2018, p. 184, paragraph 1)

The distribution of the weights was such that $w_{ij} > 0$, whatever ij.

In 2020, follow-up work provided direct support for the relation between CANAE and the cusp catastrophe model (see Appendix A in van der Maas, Dalege, & Waldorp, 2020). In this work, it was found via simulation that a CANAE network could exhibit two key features of the cusp catastrophe model (Zeeman, 1977), hysteresis and pitchfork bifurcation, under conditions with a low temperature and a distribution of weights such that $w_{ij} > 0$, whatever ij. The independent variable, the control variable in cusp catastrophe terminology, was the average value of the vertex dispositions τ_i; the dependent variable was the sum score of the system (the sum of all x_i). The system-level behavior of interest was to understand when the sum scores were at a maximum or minimum in reference to the control variable; these limiting states represent strong positive or negative attitude respectively.

Understanding the relation between CANAE and the cusp catastrophe model makes it easier to place other work in the CANAE canon, especially the theoretically oriented simulation work, all of which shares characteristics of the cusp catastrophe form–distributions of vertex baseline disposition τ_i and system temperature t were explored while fixing the distribution of weights w_{ij} so that they were virtually all positive (Dalege et al., 2018; Dalege, Borsboom, van Harreveld, Waldorp, & van der Maas, 2017; Dalege & van der Maas, 2020; van der Maas et al., 2020)[18]. The primary system behavior of interest in these simulation studies was the sum score.

So, it seems that the CANAE theory evolved quickly. The original form was a content-addressable associative memory system that assumed that vertex states represented endorsement or not of the evaluative reaction directly. Its function was to store the states to which it was exposed and recall them when cued by social context. The second form was a cusp catastrophe system in which the vertex states x_i were to be interpreted as the valence of the evaluative reaction directly; the principal system measure of interest was the sum score. Its function was to represent a strong positive or negative valenced state of the system and to demonstrate the control of the system using the distribution of the baseline dispositions τ_i over a range of values of system temperature.

Now we can attempt a sensible answer to our question. In the small world regime CANAE behaves as if it were a cusp catastrophe form. The distribution of the weights were largely positive (between 99 and 90 percent) and the system-level end state we observed was extreme: a sum score very close to 1000, equal to the number of vertices

---

[18] For some cases this assertion is based on distributions with expected values that only approximate the assertion of virtually all positive weights.

in the network. It seems, then, that the small world constraint dictated by CANAE, ceteris paribus, supports well a cusp catastrophe form of CANAE. In short, the CANAE small-world assertion implies that the function of CANAE is to provide cusp catastrophe dynamics for attitude formation and change. [19]

Critical Issues.   Our analysis of the evolution of the CANAE theory and our inevitable conclusion raised two questions in respect to both learning and measurement in CANAE. Throughout the CANAE literature, learning of the network weights w_{ij} and vertex dispositions τ_i was not addressed directly. Instead, these parameters were fixed prior to simulation[20], with an acknowledgement that Hebbian learning might be a fruitful path for future work (Dalege et al., 2018; van der Maas et al., 2020). As we describe next, Hebbian learning seems reasonable for the content-addressible memory form of CANAE but not the cusp catastrophe form. In fact, it is not clear how learning is relevant for the cusp catastrophe form. With respect to issues of measurement, we note that the original CANAE form was cast as a new psychometric measurement approach:

> ...a realistic psychometric conceptualization of attitudes (Dalege et al., 2016, pg. 3, para 4).

The CANAE canon subscribes to a statistical measurement model for generating network parameters from attitudinal survey data (van Borkulo et al. (2015) IsingFit method) for the purposes of simulation using the Ising model (Dalege et al., 2016).[21] Yet, the sole example in the CANAE literature of this measurement model applied to Ising simulation deviates significantly from the measurement model with little justification; this deviation has striking effects on the dynamics of the system as we will demonstrate shortly.

Before we address the issues of learning and measurement in detail, some observations will prove useful to keep in mind. First, both forms of CANAE are memory systems in which the weights w_{ij} and baseline dispositions τ_i of vertices store information about the world and that are responsive to cues for retrieval of an attitude. Second, the model operation of both forms of CANAE are identical; they use Ising-like or Hopfield-like dynamics. Third, the distributions of the network weights are different across forms. The original associative memory form does not constrain the weights, but learns them from data; the cusp catastrophe form of CANAE fixes the weights to be virtually all positive in value. Finally, the semantics of vertex states are distinct–in the cusp catastrophe form these represent valence directly while the original form represents endorsement or not of a belief or evaluative reaction.

To ground our comparison, we invoke the *associative memory problem*, a fundamental notion that arose from the study of neural networks in the 1980s:

---

[19] In Study 2 we set the baseline dispositions τ_i to zero; we conjecture that systematic changes in τ_i would provide the same functionality in Study 2 as does a control variable in the cusp catastrophe CANAE form.

[20] Dalege, Borsboom, van Harreveld, and van der Maas (2017) is an exception to this rule; we discuss it in detail shortly.

[21] There are some other statistical methods used in the CANAE literature but they do not address simulation of the Ising-like attitude model; instead, they are used to analyze data directly (e.g., Dalege, Borsboom, van Harreveld, Waldorp, & van der Maas, 2017; Dalege & van der Does, 2022).

> Store a set of p patterns ξ_i^μ in such a way that when presented with a new
> pattern ζ_i, the network responds by producing whichever one of the stored
> patterns most closely resembles ζ_i. (Hertz et al., 1991, p. 11, paragraph 1).

In this formulation, $\mu = 1, 2, ..., p$ represent the patterns and the vertices are represented by $i = 1, 2, ..., N$. A solution to the associative memory problem is to find the set of weights, w_{ij} and dispositions τ_i that result in this behavior; if successful, the stored patterns ξ_i^μ represent attractors. The associative memory problem provides a useful, clear basis for comparison of the two CANAE forms, especially in respect to learning.

The associative memory problem captures well the original CANAE form, as exemplified by Study 1. To recapitulate, we used the IsingFit method (van Borkulo et al., 2015) to generate Ising simulation parameters followed by the presentation of all possible patterns (the full configuration space) to the system; this method revealed the stored patterns in the system (we called these referent fixed-points in Study 1). Hebbian learning, as referenced in the 2018 theoretical paper (Dalege et al., 2018), would also have served the purpose of generating the Ising simulation parameters. In short, the associative memory problem and the original CANAE form are well aligned.

For the cusp catastrophe form of CANAE, the associative memory problem is, in a sense, inverted. We might state it like this: Given two desired patters ξ_{min}^μ and ξ_{max}^μ the problem is to understand the system's behavior in respect to these two patterns via the control variable, namely the average value of the vertex distributions $\bar{\tau}$. These two patterns are the all negative states or all positive states such that whatever i, $x_i = -1$ or $x_i = 1$; they reflect extreme negative and positive attitude, respectively. The interest is not to know which specific patterns, when presented, map to which stored attractors but to understand the partitioning of the pattern space in respect to three regions: (i) where $\bar{\tau} < \epsilon_{min}$ and the system generates ξ_{min}^μ, (ii) where $\bar{\tau} > \epsilon_{max}$ and the system generates ξ_{max}^μ, and (iii) where $\epsilon_{min} < \bar{\tau} < \epsilon_{max}$ and the system yields a mixture of system states. (The latter region is expected to be small given a low temperature.) In sum, the associative memory problem amounts to storing two patterns in such a way that the system is controlled by $\bar{\tau}$ to provide the right partitioning of the configuration space. Storage, then, becomes trivial. A simple solution is to fix all weights greater or equal to some positive value, a solution that, in practice, has support in the CANAE literature (Dalege et al., 2018; Dalege, Borsboom, van Harreveld, Waldorp, & van der Maas, 2017; Dalege & van der Maas, 2020; van der Maas et al., 2020).

Having such a simple solution readily at hand puts into question the suitability of the cusp catastrophe form of CANAE as a model of attitude learning, at least in terms of Hebbian learning. We suggest that this is a major point of exploration for future work on CANAE. Prior work on fitting cusp catastrophe models to attitudinal data may prove useful (see Van Der Maas, Kolstein, & Van Der Pligt, 2003) as might advanced optimization techniques in neural networks (see Aggarwal et al., 2018; Rojas, 2013, for efforts in this direction) .

Now we move to the issue of measurement. Earlier in this section we pointed out that the original content-addressable associative memory form of CANAE was fashioned as an alternative measurement model for attitudinal data. The first (and only) attempt at Ising simulation in conjunction with the CANAE measurement model departed from the measurement model in a somewhat peculiar manner (e.g., Dalege, Borsboom, van Harreveld, & van der Maas, 2017)[22]; we say peculiar because although it seems like a

---

[22] This effort used the IsingFit method for estimation of the network, a method that was developed for

small deviation from the measurement model, it comes with dramatic effects on the dynamics of the network, as we will show next.

The deviation worked as follows. Prior to simulation, the network was recoded such that the subset of the weights that connected vertices of opposite valence were reversed in sign (valence was judged by the researcher and was not inherent in the data or represented in the simulation). In practice, this technique had the effect of making virtually all of the network weights positive in value, an effect that is reminiscent of the cusp catastrophe form of CANAE.

The motivation for this technique was similar to that of recoding survey items in order to align the interpretation of item values. The following quote is in reference to Ising-like attitude models:

> ... the connections may not be all equal, they will be mostly positive (after rescaling). That is, we conjecture that it is generally possible to define all relevant nodes (for instance, regarding the consumption of meat) such that all positive values represent a pro attitude and all negative values represent a contra attitude. This is standard practice in the analysis of attitude questionnaires. (van der Maas et al., 2020, Appendix A, paragraph 2).

The reverse coding scheme seemed problematic to us; we conjectured that it would not preserve the dynamics of the system. In Study 1, we extended the simulation effort in question (Dalege, Borsboom, van Harreveld, & van der Maas, 2017) exactly as it was conducted and thus our findings were resultant of the reverse coding scheme (see Figure 1). Now, we will compare the recoded vs non-recoded networks to surface the difference in dynamics. Figure 17 serves as a direct comparison to Figure 1. In the former, reverse coding was not used; in the latter, it was. The principal points of comparison are when the nodes "Angry" and "Afraid" were perturbed. In the recoded system, these were very similar to all other perturbations. In the non-recoded system, these were markedly different.   Thus, our conjecture stands.

Reverse coding, in this case, generated a network with nearly all weights w_{ij} positive. We surmise that this feature of the network is the basis for the observed differences in dynamics between the recoded and non-recoded networks. An open question for future CANAE work in this vein is whether recoding, with realistic survey data, will generate similar distributions of weights. It likely will. Reverse coding will, typically, target negative correlations precisely because in an attitudinal survey instrument most negative correlations will arise from complementary valenced items; in other words, the correlations are dictated, to a large degree, by what has been described as a positive manifold of attitudinal items (Dalege et al., 2016).

Formal Predictions in Attitude Networks

Study 1 and Study 2 leaned heavily on simulation methods, a fruitful avenue towards understanding a system and its dynamics. But, when relying on simulation alone to understand a system, one runs the risk that insights obtained through example simulations are non-generic. We now turn the discussion to the value of formal mathematical analysis of attitudinal systems and their dynamics, something that applies generally to psychological networks. Formal analytical methods have the

---

estimation of Ising-like models from binary data (see van Borkulo et al., 2015).

potential to not only increase understanding but also to generate unambiguous experimental predictions. We do this by example using a prominent computational model of attitudes, the Attitudes as Constraint Satisfaction (ACS) model (Monroe & Read, 2008). In outline form, we will provide: (i) the requisite technical details of Graph Dynamical Systems (the mathematical framework we use), (ii) a high-level summary review of the ACS model and its gaps in terms of making precise testable predictions, (iii) a sketch of how we propose to move from mathematical analysis to useful experimental predictions using the ACS model (Monroe & Read, 2008).

Graph Dynamical Systems.   Although Graph Dynamical Systems (GDS) appears outside of the psychological context, it is well suited for understanding attitude dynamics; GDS was developed in the context of socio-technical systems writ large, but was meant as a general abstraction for modeling and analyzing the discrete dynamics of networked systems.

The mathematical and computational theory of GDS (see, e.g., Goles and Martinez (1990); H. S. Mortveit (2023); H. S. Mortveit and Reidys (2001, 2007); Rosenkrantz, Marathe, Hunt III, Ravi, and Stearns (2015)) is largely concerned with finite state sets such as $\{0, 1\}$ and specific update mechanisms used to assemble local dynamics on agents[23] into global dynamics of the complete system. Formally, a sequence of vertex functions $(f_i)_i$ indexed by the agents will, by applying an update scheme U, assemble to a map $F_U \colon K^n \longrightarrow K^n$ where K is the state set of each agent. For example, for a parallel update scheme with n agents/vertices, we have

$$F_U\Big(x = (x_1, \ldots, x_n)\Big) = \Big(f_1(x), \ldots, f_n(x)\Big) , \tag{1}$$

where the function f_i captures the behavior of vertex i. The variables which the functions f_i consumes capture the dependencies among the corresponding agents; we encode these through the dependency graph G. *In terms of the present article, contemporary computational models of attitudes–e.g., Hopfield models, Ising-like models, fully recurrent neural networks–are special cases of GDS.*

Existing mathematical and computational theory of GDS deals with how structural properties of the functions f_v, properties of the network G, and the choice of update mechanism translate into properties of the system, captured through the state space dynamics. All standard questions and topics from dynamical system theory such as stability and control are studied. For example, it is well known that binary threshold GDS under sequential update mechanisms (see, e.g., Goles and Olivos (1980); H. S. Mortveit and Reidys (2007)) have only fixed points as attractors and these are invariant with respect to the choice of update sequence (H. S. Mortveit & Reidys, 2007), while the parallel update method, it turns out that periodic orbits of length 2 can also manifest (Goles & Martinez, 1990).

The examples we provide below mark a way of using GDS to build a rigorous foundation of attitudinal networks. Our focus will address the Attitudes as Constraint Satisfaction Monroe and Read (2008) model. The general form of this models is captured completely in the GDS formalism:

- There is a graph $G = G(V, E)$ consisting of a collection of beliefs (vertices from a set V) and relations between them (weighted edges from a set E). (Social

---

[23] Agents map onto vertices or nodes of a graph.

psychologists will be familiar with nodes and weights (vertices and relations) in an artificial neural network.)

- Each vertex $i \in V$ is assigned a dynamic state $x_i \in K_i$ where K_i is the state set for that vertex. Generally, we have $K_i = [a_i, b_i] \subset \mathbb{R}$ where a_i and b_i are bounds on the vertex state values.

- The system state is $x = (x_1, x_2, \ldots, x_n)$.

- The edges, which are directed, are defined by a real-valued matrix $W = [w_{ij}]$. An edge e from vertex i to vertex j is written $e = (i, j)$ and has associated edge weight w_{ij}.

- For each vertex $i \in V$ there is a function $\sigma_i \colon \prod_{i=1}^n K_i \longrightarrow \mathbb{R}$ performing a local computation for vertex i that captures both vertex biases and some form of coupling with other vertices through the the matrix W, e.g., $\sigma_i(x) = \sum_{i \neq j} w_{ij} x_j$.

- Finally, for each vertex i there is a vertex function of the form $f_i = \phi_i \circ \sigma_i$ where $\phi_i \colon \mathbb{R} \longrightarrow \mathbb{R}$ is for instance a threshold function, like Heaviside.

- These kinds of models are typically explored through discrete-time, asynchronous simulations where, for each time step, one selects a vertex i and evaluates f_i, and instantiates a state change only for vertex i.

Attitudes as Constraint Satisfaction. The ACS model was developed to demonstrate, as proof-of-concept, that dynamic network models could capture some of the key empirical patterns in attitude research via simulation; no mathematical analysis was provided. This model was not designed to capture real human attitudinal contexts or to capture a specific set of experimental data. Instead, the ACS leveraged a high degree of abstraction and a low degree of specificity to offer a proof-of-concept model, one that would spur future development.

The primary measure of the ACS model (Monroe & Read, 2008) was the dynamics of a single vertex, called the evaluative vertex (x_e) that served as the evaluation of the attitude object. The graph was partitioned into two competitive substructures–one to represent knowledge related to the attitude object (e.g., the presidential candidate is kind and intelligent) and the other to capture persuasive attempts against the existing knowledge. A typical simulation trained the model to gravitate toward a positive evaluation of the attitude object, i.e., a positive value of x_e. After training, the model was probed and perturbed to test the effects, theoretically, of specific kinds of persuasion, reasoning (e.g., motivated reasoning), mere thought on polarization, and elaboration likelihood. Other measures were used to provide some rudimentary understanding of the operation of the system (e.g., the energy of the system as coherence; the average states of the vertices in different partitions).

Across a series of simulation experiments, a set of experimental factors captured aspects of the system that mapped onto real-world conditions of interest and features of variability in the structure of the system, e.g., size of knowledge structure (number of beliefs associated with the attitude object), relations of partitions (degree of competition between persuasion and knowledge), strength of the persuasion attempt (number of persuasion vertices), and finally, a form of processing capacity limitation.

The objective of the ACS model was to demonstrate that certain configurations of initial conditions (e.g., the distribution of weights in W), learning (which typically

fortified the initial conditions), and parametric configurations of the factors (e.g., size of the network, structural changes, capacity) could, in principle, mimic the coarse-grained features of key experimental phenomena.

To summarize, the set of ACS simulations were, by design, highly-idiosyncratic, post-hoc instantiations of highly-stylized constructions. Rigorous methods were not employed, nor have they been since, that would characterize this model in a systematic way. Also by design, the proof-of-concept simulations did not yield quantitative predictions that were amenable to experimental test.

From Mathematical Analysis to Experimental Predictions. In this section we will (i) provide the formal description of the specific GDS form we aim to use for the proposed demonstration, (ii) define precisely the ACS model as a GDS, (iii) provide a stylized, textbook-like example of an experimental prediction from a simple formulation of the ACS, (iv) demonstrate, by example, how we envision the formulation of theoretically important experimental predictions in the future.

A GDS we would consider for the ACS would be a *weighted, block sequential graph dynamical system* over a set $V = \{1, 2, \ldots, n\}$ that is constructed from a sequence of vertex functions $F = (f_i)_{i=1}^n$ and a map U that for each time step $t \geq 0$ assigns a subset $U(t) \subset V$ whose states are to be updated at that time. Given some initial system state $x(0)$, the dynamics of the system state $x(t) = (x_1(t), x_2(t), \ldots, x_n(t))$ is given by:

$$x_i(t+1) := \begin{cases} x_i(t) , & \text{if } i \notin U(t) \\ f_i\big(x(t)\big) , & \text{if } i \in U(t) \end{cases} \tag{2}$$

The dependency graph G associated to F has vertex set V and edges all (i, j) for which f_j depends non-trivially on x_i. The graph G captures the possible interactions among vertices. We associate to F the matrix $W = [w_{ij}] \in M_n(\mathbb{R})$ of edge weights; here it is assumed that $w_{ij} \neq 0$ if and and only if (i, j) is an edge in G.

Definition 1 *We set $V = \{1, 2, \ldots, n\}$ and specify the following:*

- *An* object *vertex $v = 1$ with state $x_1 \in \{0, 1\}$, capturing the absence/presence of an* object *to be evaluated, and an* evaluation *vertex $v = n$ with $x_n \in [-1, 1] \subset \mathbb{R}$. Here $x_n < 0$ (resp. $x_n > 0$) models a negative (resp. positive) attitude toward the object, with $|x_n|$ representing the* strength *or* degree of polarization *towards the object.*

- *A partition of the remaining vertices $\{2, 3, \ldots, n-1\}$ into non-empty subsets C and P. The set C is called the* cognitive *partition and represents* features, concepts or interior beliefs *held about the object, while P, the* persuasion *partition, represents* exterior persuasive influences *regarding the object.*

- *Parallel update: for all time steps $t \geq 0$ we have $U(t) = V$.*

- *Vertex functions: let e_1 denote the unit vector $(1, 0, \cdots, 0)$, W_i the ith row of W, and $\langle x, x' \rangle$ the inner product of vectors x and x'. The vertex functions are defined by*

$$f_1(x) = \langle e_1, x \rangle , \quad f_{1 < i < n}(x) = 1 \big/ (e^{-\langle W_i, x \rangle} + 1) , \quad \text{and,} \quad f_n(x) = (e^{\langle W_n, x \rangle} - 1) \big/ (e^{\langle W_n, x \rangle} + 1) .$$

- *Stopping criterion: for $\theta \in \mathbb{R}$ the θ-stopping time t^* is the smallest time step such that the norm $\|x(t^* + 1) - x(t*)\| \leq \theta$.*

Next we provide an example of predictions for a hypothetical experiment. We focus on what parameter regimes and weight ranges may give rise to successful or unsuccessful persuasion attempts. Although simple, this provides key elements of what we envision would apply to more advanced models addressed in the psychological literature.

We use a system with four vertices $V = \{o, p, c, e\}$, where o is the object, e the evaluation, c is for cognition, and p is for persuasion. The weight matrix W has four non-zero entries given by $w_{oc} = \alpha$, $w_{ce} = \alpha'$, $w_{oe} = \gamma$, and $w_{pc} = \delta$ as shown on the left in Figure 18. The states $x_o = 1$ and $x_c = 1$ represent the presence of the object and *cognitive engagement* with the object, while $x_e > 0$ represents a positive attitude towards the object. The α parameter represents *active attention* towards the object while δ represents a competitive coupling to an exterior *persuasive influence* represented by the state $x_p = 1$. This attention competition is modeled by tuning the value of the parameter δ from 0 to 1. Finally, α' represents the *cognitive contribution* to the attitude value of e while $\gamma > 0$ represents a *automatic associative bias* towards the object. As the δ-tuning takes place, we wish to study which scenarios (parameter settings) lead to *compliance with persuasion* with the persuasive influence which seeks to change the $x_e > 0$ (positive attitude) to $x_e < 0$ (negative attitude).

In our case, the vertex functions of the ACS model for vertices c and e are given by $f_c(x) = \alpha x_o + \delta x_p$, and $f_e(x) = \alpha' x_c + \gamma x_o$ with the remaining two being constant functions $f_o(x) = 1$ and $f_p(x) = 1$. We note that any initial state $x(0) = (x_o = 1, x_e = 1, x_c = 1, x_p = 1)$ is eventually mapped onto the fixed point

$$x_o = 1, \quad x_p = 1, \quad x_c = \alpha + \delta, \quad \text{and} \quad x_e = \alpha'(\alpha + \delta) + \gamma \, . \tag{3}$$

The boundary in parameter space separating successful and unsuccessful persuasion can be obtained as the manifold defined by equating x_e in Equation (3) to 0, that is, $\gamma = -\alpha'(\alpha + \delta)$. *Here is the key insight: the expression for x_e allows one to (a) identify which parameters to target in an experiment, and to (b) quantify the magnitude of adjustments to the chosen parameter(s) in order to obtain a specific outcome.* With a model having many parameters, one may want to restrict this space by introducing relations among them. In this example, we relate α' and δ through the function f as $\alpha' = f(-\delta^2)$. If we control x_c to be 1 (via $\alpha + \gamma = 1$), we can derive $\gamma = \delta^2$ to understand the relation between the degree of persuasion and the degree of automatic associative bias. Figure 19 illustrates an experimental predictions in terms of when a persuasion attempt would be successful or not. In practice, one would relate parameters and possible constraints to experimental mechanisms and controls.

It is important to revisit the purpose of this exercise, a simple, textbook-like example of the process from mathematical formulation to experimental prediction in attitudinal networks. It was not to show the nature of the kinds of predictions we envision for future work (we discuss this below) but to show what we mean by making precise empirical predictions on dynamic networks.

Next we provide a sketch of what formal methods might look like in a more realistic setting using the ACS model. We focus on a question of direct interest to attitudes. This question has two parts and can be broadly categorized as having directly to do with the act of persuasion: *What is the ratio of persuasion elements to cognitive elements for a successful persuasion, and is this ratio an invariant with respect to any characteristics of the act of persuasion?*

To formalize this (using some reassignment of parameters in comparison to the above example), we let $M = (W, F, U)$ be the GDS formulation of the ACS model previously described. The ratio of interest, call it Q, then naturally maps in our formulation to

$$Q(M) := \frac{|C|}{|P|}.$$

We can control for the size of the total system, which means fixing $n = \dim(W)$. We let $|P| := \alpha(n-2)$ where $\alpha \in [0, 1]$ represents the fraction of nodes in the system that are persuasive. We label $M_\alpha := (W(\alpha), F(\alpha), U(\alpha))$ to denote this interpretation. Since $|C| + |P| = n - 2$ we have in terms of α

$$Q(M_\alpha) := \frac{n-2}{|P|} - 1 = \frac{n-2}{\alpha(n-2)} - 1 = \frac{1-\alpha}{\alpha}.$$

Thus, for a system of fixed size n, any $x(0) \in \mathbb{R}^n$ initial state now has an interpretation of an α-fraction of its coordinates belonging to persuasive elements. We denote $x(t_\alpha^*) := M_\alpha[x(0)]$ the state of the system M_α at its θ-stopping time t_α^*, for the fixed initial condition $x(0)$. We can then propose to study variational statements of the form

$$|\alpha - \alpha'| \le \epsilon, \quad \implies \quad |x_n(t_\alpha^*) - x_n(t_{\alpha'}^*)| \le \delta,$$

or more generally ,

$$|\alpha - \alpha'| \le \epsilon, \quad \implies \quad ||x(t_\alpha^*) - x(t_{\alpha'}^*)|| \le \delta,$$

where $|| \cdot ||$ denotes suitably chosen matrix/vector norms, and $\delta \in \mathbb{R}$ represent the errors in evaluation and in terminal system state respectively, both taken under α-perturbations controlled by $\epsilon \in \mathbb{R}$.

The implications formulated above have natural interpretations which will be of interest. For instance, a parameter pair (ϵ, δ) satisfying the first implication says that a variation of at most ϵ from the current persuasive strength α, would guarantee a change of no more than δ in the attitude the system exhibits under the constant initial circumstances of $x(0)$. Namely, all things being equal, we can relax/increase our persuasion by no more than ϵ and expect a drift of no more than δ in attitude.

We formulate the question in this manner as it conveniently allows us to see the problem in terms of a continuous parameter α (a proxy for the proportion of persuasive nodes). Using this formulation this naturally becomes a question about the continuity of the dynamics of the system when α varies. For instance for a fixed α, if the collection of vertex local functions is interpreted as a phase space function that is iterated under block sequential update, and the local vertex functions happen to be such that F is contractive, then this becomes an iterated function system. Then, by Hutchinson (1981) we know a unique fixed set exists for such a system which is now α-parametrized. The "volumes" of these sets then bound the attitude changes possible for any initial condition state chosen from such a set, and this bound can change as α is varied.

This more realistic example provides a sketch of one way to expand formal methods to more realistic psychological networks. The key takeaway point is that, by example, persuasion attempts may be modeled by the ratio of cognitive vertices to persuasion vertices. Using this formulation, we could consider variational methods to make experimental predictions about the degree of change in the attitude network given

a specific degree of persuasion.

Using variational and other formal methods, it may be feasible to develop very nuanced hypotheses related to a broad set of research questions per the ACS:

- Is it possible to identify vertices that would be more effective for persuasion attempts. For example, assume we are limited to only perturbing 10% of the persuasion vertices, under what conditions should we (should we not) target specific vertices. Assuming such conditions do exist, what defines the heterogeneity in effectiveness across vertices (e.g., centrality of vertices)?

- Assume some vertices of the cognition partition are not measured. How does this affect our understanding of persuasion and learning?

- Assume that the set of vertices that are measured includes both relevant and irrelevant vertices in the cognition partition. How does this affect our understanding of persuasion and learning? Can we model the irrelevant vertices as noise?

- What is the effect of precision in estimation of edge values on the dynamics of the system? Are binary values sufficient, especially as networks get bigger? Or do we need more precise/granular measurement? What level of precision do we need?

- How do we represent/capture learning in the model? Change in vertex strength? Change in edge strength?

- Can we identify points of potential maximal change or vulnerability using the wide range of network characteristics that have been developed. How would changes in those characteristics influence dynamics?

In summary, we described a mathematical approach to understanding dynamics on graphs, graph dynamical systems, that hitherto has never been applied to attitude networks or psychological networks. We provided two examples in an effort to illustrate the value of this approach for both understanding attitudinal systems but also for making precise predictions. We surmise that such an approach might help to address the dynamics issue outlined in recent work in clinical psychology (see Bringmann, 2021; Bringmann et al., 2019; Bringmann & Eronen, 2018; Burger et al., 2020; A. O. Cramer et al., 2016; Haslbeck et al., 2021; Wichers et al., 2015). In short, formal methods for dynamical systems on graphs are mature and apt for the problem of attitude modeling and, more generally, the problem of dynamics on psychological networks.

Conclusions

We conclude with the assertion that the two claims of CANAE addressed in this article were not supported. The effects of perturbing an evaluative reaction was demonstrated to not be a function of its network centrality (Study 1). The hypothesis that a small-world network structure offers a trade-off between attitudinal consistency and accuracy was not the right characterization (Study 2); the trade-off is between consistency and capacity of the system. The small-world does provide a trade-off, but at the expense of very limited capacity of the system. The results of Study 2 spurred us to dig deep into the evolution of CANAE, looking for hints of the purpose of such a limited capacity system in terms of attitudes. We found that there are two distinct

forms of CANAE; one is a content-addressable associative memory system, as described in the introduction and the second is a network variant of the cusp catastrophe model. The differences in these two forms are significant enough to warrant divergent future research directions.

Our assertion is qualified by the fact that our work was based on simulation work that was not supported by formal methods. We offer in the General Discussion a glimpse of what we mean in terms of formal methods both towards deeper understanding of psychological network dynamics and for making more rigorous theoretical predictions of such.

Author Note

References

Aggarwal, C. C., et al. (2018). Neural networks and deep learning. *Springer*, *10*(978), 3.

Anderson, J. A., & Rosenfeld, E. (1989). *Neurocomputing: Foundations of research.* MIT Press.

Borsboom, D. (2008). Psychometric perspectives on diagnostic systems. *Journal of Clinical Psychology*, *64*(9), 1089–1108. doi: 10.1002/jclp.20503

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, *16*(1), 5–13. doi: 10.1002/wps.20375

Borsboom, D., & Cramer, A. O. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, *9*(1), 91–121. doi: 10.1146/annurev-clinpsy-050212-185608

Bringmann, L. F. (2021, Oct). Person-specific networks in psychopathology: Past, present, and future. *Current Opinion in Psychology*, *41*, 59–64. doi: 10.1016/j.copsyc.2021.03.004

Bringmann, L. F., Elmer, T., Epskamp, S., Krause, R. W., Schoch, D., Wichers, M., . . . Snippe, E. (2019). What do centrality measures measure in psychological networks? *Journal of Abnormal Psychology*, *128*(8), 892.

Bringmann, L. F., & Eronen, M. I. (2018). Don't blame the model: Reconsidering the network approach to psychopathology. *Psychological review*, *125*(4), 606.

Brush, S. G. (1967, Oct). History of the lenz-ising model. *Reviews of Modern Physics*, *39*(4), 883–893. doi: 10.1103/RevModPhys.39.883

Burger, J., van der Veen, D. C., Robinaugh, D. J., Quax, R., Riese, H., Schoevers, R. A., & Epskamp, S. (2020). Bridging the gap between complexity science and clinical practice by formalizing idiographic theories: a computational model of functional analysis. *BMC medicine*, *18*(1), 1–18.

Chambon, M., Dalege, J., Elberse, J. E., & van Harreveld, F. (2022). A psychological network approach to attitudes and preventive behaviors during pandemics: A covid-19 study in the united kingdom and the netherlands. *Social Psychology and Personality Science*, *13*(1), 233-245.

Cipra, B. A. (1987, Dec). An introduction to the ising model. *The American Mathematical Monthly*, *94*(10), 937–959. doi: 10.1080/00029890.1987.12000742

Conrey, F. R., & Smith, E. R. (2007). Attitude representation: Attitudes as patterns in a distributed, connectionist representational system. *Social Cognition*, *25*(5), 718–735.

Cramer, A. O., Van Borkulo, C. D., Giltay, E. J., Van Der Maas, H. L., Kendler, K. S., Scheffer, M., & Borsboom, D. (2016). Major depression as a complex dynamic system. *PloS one*, *11*(12), e0167490.

Cramer, A. O. J., Waldorp, L. J., Maas, H. L. J. v. d., & Borsboom, D. (2010, Jun). Comorbidity: A network perspective. *Behavioral and Brain Sciences*, *33*(2–3), 137–150. doi: 10.1017/S0140525X09991567

Dalege, J., Borsboom, D., van Harreveld, F., van den Berg, H., Conner, M., & van der Maas, H. L. (2016). Toward a formalized account of attitudes: The causal attitude network (can) model. *Psychological review*, *123*(1), 2.

Dalege, J., Borsboom, D., van Harreveld, F., & van der Maas, H. J. L. (2017). Network analysis on attitudes: A brief tutorial. *Social Psychological and Personality Science*, *8*(5), 528-537.

Dalege, J., Borsboom, D., van Harreveld, F., & van der Maas, H. L. J. (2018, Oct). The attitudinal entropy (ae) framework as a general theory of individual attitudes. *Psychological Inquiry*, *29*(4), 175–193. doi: 10.1080/1047840X.2018.1537246

Dalege, J., Borsboom, D., van Harreveld, F., & van der Maas, H. L. J. (2019). A network perspective on attitude strength: Testing the connectivity hypothesis. *Social Psychological and Personality Science*, *10*(6), 746-756.

Dalege, J., Borsboom, D., van Harreveld, F., Waldorp, L. J., & van der Maas, H. L. J. (2017, Jul). Network structure explains the impact of attitudes on voting decisions. *Scientific Reports*, *7*(1), 4909. doi: 10.1038/s41598-017-05048-y

Dalege, J., & van der Does, T. (2022). Using a cognitive network model of moral and social beliefs to explain belief change. *Science Advances*, *8*(0137), 1-15.

Dalege, J., & van der Maas, H. L. J. (2020, Nov). Accurate by being noisy: A formal network model of implicit measures of attitudes. *Social Cognition*, *38*(Supplement), s26–s41. doi: 10.1521/soco.2020.38.supp.s26

Ehret, P. J., Monroe, B. M., & Read, S. J. (2015). Modeling the dynamics of evaluation: A multilevel neural network implementation of the iterative reprocessing model. *Personality and Social Psychology Review*, *19*(2), 148–176.

Fried, E. I. (2020, October). Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry*, *31*(4), 271–288. doi: 10.1080/1047840X.2020.1853461

Gerstner, W., Kistler, W. M., Naud, R., & Paninski, L. (2014). *Neuronal dynamics: From single neurons to networks and models of cognition.* Cambridge University Press.

Goles, E., & Martinez, S. (1990). *Neural and automata networks: Dynamical behaviour and applications.* Kluwer Academic Publishers.

Goles, E., & Olivos, J. (1980). Periodic behavior in generalized threshold functions. *Discrete Mathematics*, *30*, 187–189.

Haslbeck, J., Ryan, O., Robinaugh, D. J., Waldorp, L. J., & Borsboom, D. (2021). Modeling psychopathology: From data models to formal theories. *Psychological Methods*.

Hertz, J., Krogh, A., & Palmer, R. G. (1991). *Introduction to the theory of neural computation.* Addison-Wesley.

Hinton, G. E., & Anderson, J. A. (1989). *Parallel models of associative memory* (Updated ed.). Lawrence Erlbaum Associates.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*, 2554–2558.

Hutchinson, J. (1981). Fractals and self-similarity. *Indiana Univ. Math. J.*, *30*, 713–747.

Liu, J. H., & Latané, B. (1998). The catastrophic link between the importance and extremity of political attitudes. *Political Behavior*, *20*, 105–126.

Monroe, B. M., & Read, S. J. (2008). A general connectionist model of attitude structure and change: The acs (attitudes as constraint satisfaction) model. *Psychological Review*, *115*, 733–759. doi: 10.1037/0033-295X.115.3.733

Mortveit, H., & Reidys, C. (2007). *An introduction to sequential dynamical systems.* Springer Science & Business Media.

Mortveit, H. S. (2023). Asynchronous, finite dynamical systems. *Natural Computing*, *22*, 357–377. doi: 10.1007/s11047-023-09944-3

Mortveit, H. S., & Reidys, C. M. (2001). Discrete, sequential dynamical systems. *Discrete Mathematics*, *226*, 281–295. doi: 10.1016/S0012-365X(00)00115-1

Mortveit, H. S., & Reidys, C. M. (2007). *An introduction to sequential dynamical systems.* Springer Verlag. doi: 10.1007/978-0-387-49879-9

Muthukrishna, M., & Henrich, J. (2019, February). A problem in theory. *Nature Human Behaviour*, *3*(3), 221–229. doi: 10.1038/s41562-018-0522-1

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., . . . Vazire, S. (2022, January). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*(Volume 73, 2022), 719–748. doi: 10.1146/annurev-psych-020821-114157

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*, 1596-1618.

Orr, M., Mortveit, H. S., Lebiere, C., & Pirolli, P. (2023). A 10-year prospectus for mathematical epidemiology. *Frontiers in Psychology*, *14*. Retrieved from `https://www.frontiersin.org/articles/10.3389/fpsyg.2023.986289`

Orr, M. G., & Chen, D. (2017). Computational models in social psychology. In R. Vallacher, A. Nowak, & S. Read (Eds.), (chap. Computational Models of Health Behavior). New York: Pyschology Press/Routledge.

Orr, M. G., & Plaut, D. C. (2014). Complex systems and health behavior change: insights from cognitive science. *American journal of health behavior*, *38*(3), 404–413.

Orr, M. G., Thrush, R., & Plaut, D. C. (2013, May). The theory of reasoned action as parallel constraint satisfaction: Towards a dynamic computational model of health behavior. *PLOS ONE*, *8*(5), e62490. doi: 10.1371/journal.pone.0062490

Orr, M. G., Zeimer, K., & Chen, D. (2017). Systems science and population health. In S. Galea & A. El-Sayed (Eds.), (chap. Systems of Behavior and Population Health). Oxford: Oxford University Press.

Overwalle, F., & Siebler, F. (n.d.). A connectionist model of attitude formation and change. *Personality and Social Psychology Review*, *9*, 231–274.

Overwalle, F. V. (2007). *Social connectionism: A reader and handbook for simulations.* Psychology Press. doi: 10.4324/9780203783115

Read, S. J., & Miller, L. C. (1998). *Connectionist models of social reasoning and social behavior.* Lawrence Erlbaum Associates Publishers.

Read, S. J., Vanman, E. J., & Miller, L. C. (1997, Jan). Connectionism, parallel constraint satisfaction processes, and gestalt principles: (re)introducing cognitive dynamics to social psychology. *Personality and Social Psychology Review*, *1*(1), 26–53.

Robinaugh, D. J., Haslbeck, J. M. B., Ryan, O., Fried, E. I., & Waldorp, L. J. (2019). Invisible hands and fine calipers: A call to use formal theory as a toolkit for theory construction. *Perspectives on Psychological Science*, *16*(4), 725–743.

Rojas, R. (2013). *Neural networks: a systematic introduction.* Springer Science & Business Media.

Rosenkrantz, D. J., Marathe, M. V., Hunt III, H. B., Ravi, S., & Stearns, R. E. (2015). Analysis problems for graphical dynamical systems: A unified approach through graph predicates. In *Proceedings of the 2015 international conference on autonomous agents and multiagent systems pages* (pp. 1501–1509). Retrieved from `https://www.ifaamas.org/Proceedings/aamas2015/aamas/p1501.pdf`

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. (1986). Sequential

thought processes in pdp models. *Parallel distributed processing: explorations in the microstructures of cognition*, *2*, 3–57.

Schlegelmilch, J., & Carlin, E. (2023). *Catastrophic incentives: Why our approaches to disasters keep falling short.* Columbia University Press.

Shultz, T. R., & Lepper, M. R. (1996). Cognitive dissonance reduction as constraint satisfaction. *Psychological Review*, *103*(2), 219-240.

Simon, D., & Holyoak, K. J. (2002). Structural dynamics of cognition: From consistency theories to constraint satisfaction. *Personality and social psychology review*, *6*(4), 283–294.

Smaldino, P. E. (2020, October). How to build a strong theoretical foundation. *Psychological Inquiry*, *31*(4), 297–301. doi: 10.1080/1047840X.2020.1853463

Smith, E. R. (1996). What do connectionism and social psychology offer each other? *Journal of personality and social psychology*, *70*, 893–912. doi: 10.1037/0022-3514.70.5.893

Thompson, A. (2023). *What makes people act on climate change, according to behavioral science.* Retrieved from `https://www.scientificamerican.com/article/what-makes-people-act-on-climate-chang`

Trappenberg, T. P. (2010). *Fundamentals of computational neuroscience* (2nd ed.). Oxford University Press.

Vallacher, R. R., Read, S. J., & Nowak, A. (2017). *Computational social psychology.* Routledge. doi: 10.4324/9781315173726

van Borkulo, C. D., Borsboom, D., Epskamp, S., Blanken, T. F., Boschloo, L., Schoevers, R. A., & Waldorp, L. J. (2015, May). A new method for constructing networks from binary data. *Scientific Reports*, *4*(1), 5918. doi: 10.1038/srep05918

van der Maas, H. L. J., Dalege, J., & Waldorp, L. (2020). The polarization within and across individuals: the hierarical ising opinion model. *Journal of Complex Networks*, *2*, 1-23.

Van Der Maas, H. L. J., Kolstein, R., & Van Der Pligt, J. (2003, November). Sudden transitions in attitudes. *Sociological Methods & Research*, *32*(2), 125–152. doi: 10.1177/0049124103253773

Van Overwalle, F., & Siebler, F. (2005). A connectionist model of attitude formation and change. *Personality and Social Psychology Review*, *9*(3), 231–274.

Watts, D., & Strogatz, S. (1998). Strogatz—small world network nature. *Nature*, *393*, 440–442.

Wichers, M., Wigman, J. T. W., & Myin-Germeys, I. (2015, Oct). Micro-level affect dynamics in psychopathology viewed from complex dynamical system theory. *Emotion Review*, *7*(4), 362–367. doi: 10.1177/1754073915590623

Zeeman, E. C. (1977). *Catastrophe theory: Selected papers 1972–1977.* Addison-Wesley.

Zwicker, M. V., Nohlen, H. U., Dalege, J., Gruter, G. M., & van Harreveld, F. (2020). Applying an attitude network approach to consumer behaviour towards plastic. *Journal of Environmental Psychology*, *69*, 1-14.

	Prop. Impossible			Mean Energy			Network Properties	
	Ones	Random	Zeros	Ones	Random	Zeros	No. Flip Nodes	Prop. Edges Neg.
A	1.00	1.00	1.00	-23.14	-22.73	-23.97	0	0.00
B	1.00	1.00	1.00	-15.41	-14.96	-15.97	0	0.17
C	0.46	0.13	0.16	-07.72	-07.68	-07.99	2	0.34
D	0.04	0.00	0.00	-05.49	-05.61	-05.99	6	0.50

Table 1

Results of the Pseudo-Necker Cube Simulations.

Figure 1. Distributions of sum scores over 1000 runs of an Ising simulation (per condition). The top-left panel depicts the baseline, no perturbation condition; the remaining panels illustrate the effects of perturbation for each vertex separately. Panel labels (bold, top-center) capture the central construct meaning associated with each vertex. The top number within each panel is the Wasserstein distance of the distribution in comparison to the baseline distribution; the bottom number is the mean sum score of the distribution.

Figure 2. Relationship between vertex centrality and extent of effect. The x-axes represent vertex centrality (strength, betweenness and closeness as used in Dalege, Borsboom, van Harreveld, and van der Maas (2017)); the y-axis represents extent of effect (as either Wasserstein distance or mean of the sum scores). The cross points represent the vertices with the highest strength and betweenness.

Figure 3. The 15 point-attractors of the 1984 Reagan CANAE attitude network. The x-axis shows each ordered index of the state vector. The y-axis identifies the fixed-points by their decimal value; the order from top to bottom is by increasing frequency. States 15, 17, 19, 21 represent the negative valence nodes.

Figure 4. The relative frequency and Hamming distance (to the most frequent point attractor) of each point-attractor. The most frequent point-attractor had a frequency of 2,148,643; in other words, about half of the 2^{22} initial states were attracted to this point-attractor.

Figure 5. The relative frequency and energy of each point-attractor. The most frequent point-attractor in Figure 4 was the same as that with the lowest energy.

Figure 6. Frequency distributions of Set 1 simulations.

Figure 7. Frequency distributions of Set 2 simulations.

Figure 8. E_k for Set 1 simulations.

Figure 9. E_k for Set 2 simulations.

Figure 10. H_k for Set 1 simulations.

Figure 11. H_k for Set 2 simulations.

Figure 12. Pseudo-Necker Cube Simulation Systems in Study 2.

Figure 13. Set 3 Simulation Results: Energy of each level of rewire.

Figure 14. Set 3 Simulation Results: Endstates of each simulation.

Figure 15. Set 4 Simulation Results: Energy of each level of rewire.

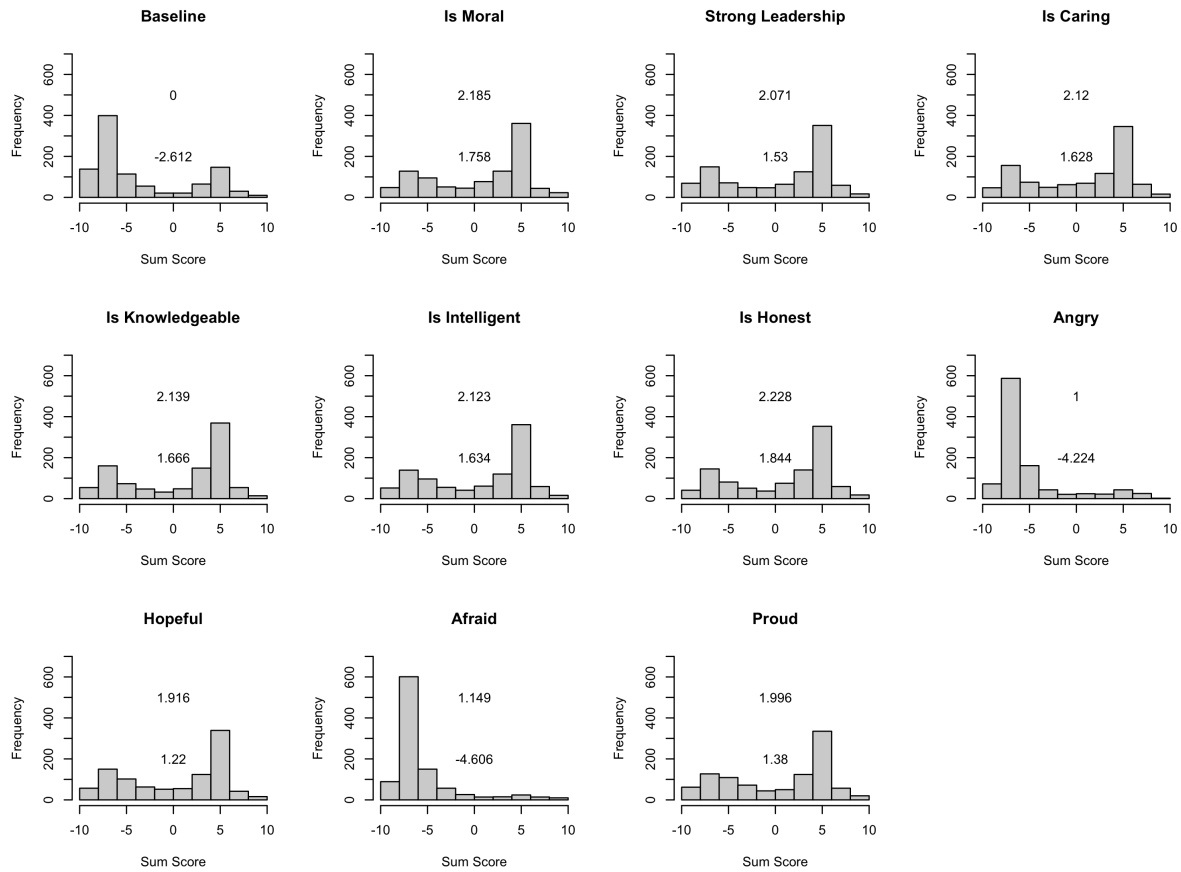*Figure 16*. Set 4 Simulation Results: Endstates of each simulation.

Figure 17. Distributions of sum scores over 1000 runs of an Ising simulation (per condition). The top-left panel depicts the baseline, no perturbation condition; the remaining panels illustrate the effects of perturbation for each vertex separately. Panel labels (bold, top-center) capture the central construct meaning associated with each vertex. The top number within each panel is the Wasserstein distance of the distribution in comparison to the baseline distribution; the bottom number is the mean sum score of the distribution. The underlying model serves as a comparison to Figure 1 in which reverse coding was used (see text for details).
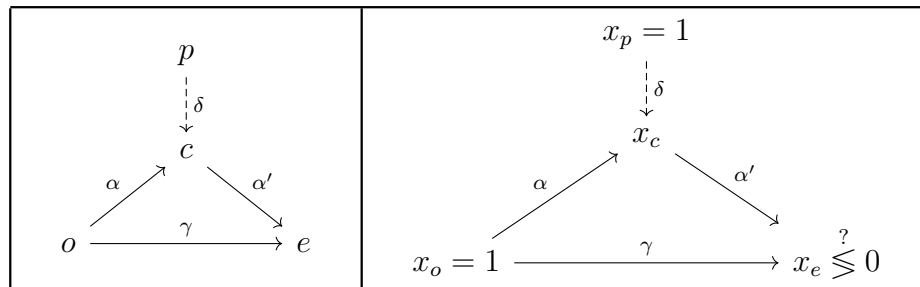
Figure 18. Left: the network of the basic example. Right: dynamics evolving over the network for state $(x_o = 1, x_p = 1, x_c, x_e)$ and assessing whether parameter choices cause compliant or non-compliant behaviors with the persuasion attempt ($x_e < 0$ or $x_e > 0$).
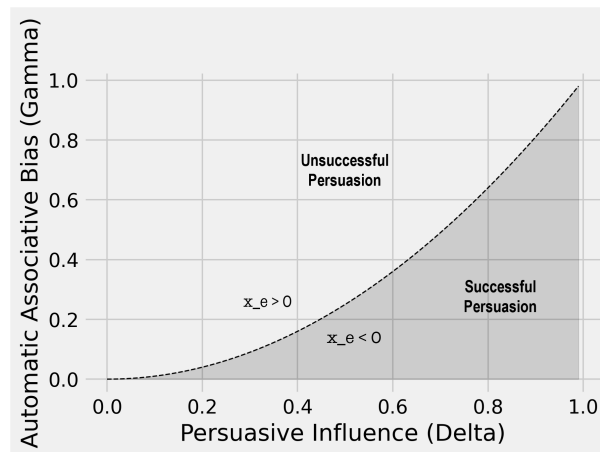

Figure 19. A simple, textbook-like example of experimental predictions via mathematical analysis for the Attitudes as Constraint Satisfaction (ACS) model (Monroe & Read, 2008). We have introduce the two parameter relation $\alpha' = f - \delta^2$ and assert that $\alpha + \delta = 1$. Under this particular choice, we obtain the boundary curve $\gamma = \delta^2$ separating the non persuadable (light; $x_e < 0$)) and the persuadable regions (dark gray; $x_e > 0$) as a function of manipulations of persuasive influence *delta* and automatic associative bias *gamma*. See text for details.

Appendix

Procedure for Computing $dist_s$ and s

(**textitNote:** In this appendix, we denote individual nodes in a network by j; in the manuscript this is not necessarily the case.)

The goal of this appendix is to elucidate the computation of the extent of the effect perturbing a single node has on the dynamics of the system. To this end, we consider our unperturbed system of 22 nodes described in Study 1. This system exhibits a set of 15 referent fixed points denoted here as a collection of 22-dimensional, $0, 1$-vectors,

$$F^R := \{f_i^r \in \{0,1\}^{22} | i = 1, \cdots, 15\}.$$

Associated to each reference fixed point f_i^r is $\nu(f_i^r) > 0$, its absolute frequency of appearance among the 2^{22} simulations conducted in Study 1 - one for each initial condition possible.

Now, for each node j of the system, we run 2^{22} simulations (one for each of the possible initial conditions), with the caveat that throughout each run, the value of node j is kept at value 1 until the system relaxes to a fixed point. We record the obtained (node specific) fixed points $f_k^j \in F^j$, and their absolute frequencies of appearance $\nu(f_k^j)$. For each such fixed point we compute a Hamming distance vector

$$v_k^j := (H(f_1^r, f_k^j), H(f_2^r, f_k^j), \cdots, H(f_{15}^r, f_k^j)).$$

We denote by $\tau_k^j := \{\kappa | H(f_i^r, f_k^j) = \min v_k^j\}$ the set of indices where the minimum Hamming distance is achieved.

Next, we create a matrix $M^j = [m_{\alpha\beta}^j]$, where $\alpha \in \{1, \cdots, |F^R|\}$ and $\beta \in \{1, \cdots, |F^j|\}$, with

$$m_{\alpha,\beta}^j := \begin{cases} \frac{\nu(f_\beta^j)}{|\tau_\beta^j|} & , \alpha \in \tau_\beta^j \\ 0 & , \alpha \notin \tau_\beta^j. \end{cases}$$

We construct the normalized vector

$$s^j := \frac{1}{\sum_k \nu(f_k^j)} [\sum_\beta M^j] = \frac{1}{\sum_k \nu(f_k^j)} [\sum_\beta m_{\alpha\beta}^j],$$

and similarly, from the reference points' absolute frequencies, we construct the normalized vector

$$r := \frac{1}{\sum_i \nu(f_i^r)} [\nu(f_\alpha^r)].$$

Finally, the extent of effect of perturbing node j is then defined as

$$E^j := ||r - s^j||_2.$$