Good afternoon. In respect for everybody’s time today, let’s go ahead and get started. So today, it is my pleasure to introduce Dr. Alexander Strang. Dr. Strang earned his bachelor’s in mathematics, in physics, as well as his PhD in applied mathematics from Case Western Reserve University in Cleveland, Ohio. Born in Ohio, so representing. He studies variational inference problems, noise propagation in biological networks, self organizing edge flows, and functional form game theory at the University of Chicago, where he is a William H. Kruskal Instructor of physics and applied mathematics. Today, he’s going to talk to us about motivic expansion of global information flow in spike train data. Let’s welcome our speaker. Okay, thank you very much. Thank you, first, for the kind invite, and for the opportunity to speak here in your seminar. So, I’d like to start with some acknowledgements. This is very much work in progress. Part of what I’m going to be showing you today is really the work of a Master’s student.
that I’ve been working with this summer, that’s Bowen, and really, I’d like to thank Bowen for a lot of the simulation, and a lot of the TE calculation I’ll show you later. This project, more generally, was born out of conversations with Brent Doiron and Lek-Heng Lim here at Chicago. Brent really was the inspiration for starting to venture into computational neuroscience. My background was much more inspired by Lek-Heng’s work in computational technology, and some of what I’ll be presenting today is really inspired by conversations with him. So, let’s start with some introduction and motivation. The motivation, personally, for this talk. So it goes back, really, to work that I started when I was a graduate student. I’ve had this long standing interest in the interplay between structure and dynamics, in particular in networks. I’ve been interested in questions like
how does the structure of a network determine dynamics of processes on that network. And, in turn, how do processes on that network give rise to structure? On the biological side...

On the biological side, in today’s talk, I’m going to be focusing on applications of this question within neural networks. And I think that this world of computational neuroscience is really exciting if you’re interested in this interplay between structure and dynamics, because neural networks encode, transmit, and process information via dynamical processes.

For example, the process, the dynamical process of a neural network is directed by the wiring patterns, by the structure of that network, and moreover, if you’re talking about some sort of learning process, then those wiring patterns can change and adapt during the learning process, so they are themselves dynamic.

In this area, I’ve been interested in questions like, how is the flow of information governed by the wiring pattern, how do patterns of information flow present themselves
in data, and can they be inferred from that data,
and what types of wiring patterns might develop during learning.
Answering questions of this type requires a couple of things.
So the very, very big picture requires a language for describing structures and patterns,
it requires having a dynamical process,
some sort of model for the neural net, and it requires a generating model that generates initial structure,
and links structure to dynamics.
Alternatively, if we don’t generate things using a model,
if we have some sort of observable or data,
then we can try to work in the other direction and go from dynamics to structure.
Today, during this talk, I’m going to be focusing really on this first piece,
on a language for describing structures and patterns,
and on the second piece,
on an observable that I’ve been working on trying to compute to use to try to connect these three points together.
So, to get started, a little bit of biology.
Really, I was inspired in this project by a paper from Keiji Miura.
He was looking at a coupled oscillator model,
this was a Kuramoto model for neural activity.
where the firing dynamics interact with the wiring.
So the wiring that couples the oscillators would adapt on a slower timescale than the oscillators themselves, and that adaptation could represent different types of learning processes. For example, the fire-together wire-together rules, so Hebbian learning, you could look at causal learning rules, or anti-Hebbian learning rules. This is just an example of one, of the system. This system of (indistinct) is sort of interesting because it can generate all sorts of different patterns. You can see synchronized firing, you can see traveling waves, you can see chaos. These occur at different critical boundaries. So you can see phase transmissions when you have large collections of these oscillators. And depending on how they’re coupled together, it behaves differently. In particular, what’s interesting here is that starting from some random seed topology, the dynamics play forward from that initial condition, and that random seed topology produces an ensemble of wiring patterns that are themselves random.
And we can think of that ensemble of wiring patterns as being chaotic realizations of some random initialization.

That said, you can also observe structures within the systems of coupled oscillators. So you can see large scale cyclic structures representing organized causal firing patterns in certain regimes.

This is a nice example where graph structure and learning parameters can determine dynamics, and in turn, those dynamics can determine structure.

On the other side, you can also think about a data-driven side instead of a model-driven side. If we empirically observe sample trajectories of some observables, for example, neuron recordings, then we might hope to infer something about the connectivity that generates them.

And so here, instead of starting by posing a model, and then simulating it and setting up how it behaves, we can instead study data, or try to study structure in data. Often, that data comes in the form of covariance matrices representing firing rates.

And these covariance matrices...
may be auto-covariance matrices with some sort of time-lag.

Here, there are a couple of standard structural approaches.

So there’s a motivic expansion approach.

This was at least introduced by Brent Doiron’s lab,

with his student, Gabe Ocker.

Here, the idea is that you define some graph motifs,

and then you can study the dynamics

in terms of those graph motifs.

For example, if you build a power series in those motifs,

then you can try to represent your covariance matrices

in terms of that power series.

And this is something we’re gonna talk about

quite a bit today.

This really, part of why I was inspired by this work is

I had been working separately

on the idea of looking at covariance matrices

in terms of these power series expansions.

This is also connected to topological data analysis,

and this is where the conversations with Lek-Heng

played a role in this work.

Topological data analysis aims to construct graphs

representing dynamical systems.

For example, you might look at the dynamical similarity
of firing patterns of certain neurons, and then try to study the topology of those graphs. Again, this leads to similar questions, but we could be a little bit more precise here for thinking in neuroscience. We can say more precisely, for example, how is information processing and transfer represented, both in these covariance matrices and the structures that we hope to extract from them? In particular, can we try and infer causality from firing patterns? And this is fundamentally an information theoretic question. And really, we’re asking, can we study the directed exchange of information from trajectories? Here, one approach, I mean, in some sense, you can never tell causality without some underlying model, without some underlying understanding and mechanism, so if all we can do is observe, then we need to define what we mean by causality. A reasonable standard definition here is Wiener causality, which says that two time series share a causal relation, so we say x causes y, if the history of x informs the future of y.
And note that here, cause, I put in quotes, really means forecasts. It means that the past, or the present of x, carries relevant information about the future of y. A natural measure of that information is transfer entropy. Transfer entropy was introduced by Schriefers in 2000, and is the expected KL divergence between the distribution of the future of y given the history of x, and the marginal distribution of the future of y. So essentially, it’s how much predictive information does x carry about y. This is a nice quantity for a couple of reasons. First, it’s zero when two trajectories are independent. Second, since it’s just defining some of these conditional distributions, it’s model free, so I put here no with a star, because generative assumptions actually do matter when you go to try and compute it, but in principle, it’s defined independent of the model. Again, unlike some other effective causality measures, it doesn’t require introducing some time-lag to define. It’s a naturally directed quantity.
We can say that the future of y conditioned on the past of x... That transfer entropy is defined in terms of the future of y conditioned on the past of x and y. And that quantity is directed, because reversing x and y does not symmetrically change the statement. This is different than quantities like mutual information or correlation, that are also often used to try and measure effective connectivity in networks which are fundamentally symmetric quantities. Transfer entropy is also less corrupted by measurement noise, linear mixing of signals, or shared coupling to external sources. Lastly, and maybe most interestingly, if we think in terms of correlations, correlations are really moments, right, second order moments. Transfer entropies, these are about entropies, these are logs of distributions, and so they depend on the full shape of these distributions. Which means that transfer entropy can capture coupling that is maybe not apparent, or not obvious just looking at second order moment type analysis. So transfer entropy has been applied pretty broadly.
It's been applied to spiking cortical networks and calcium imaging, to MEG data in motor tasks and auditory discrimination, it's been applied to emotion recognition, precious metal prices and multivariate time series forecasting, and more recently, to accelerate learning in different artificial neural nets.

For this part of the talk, I'd like to focus really on two questions. First, how do we compute transfer entropy, and then second, if we could compute transfer entropy, and build a graph out of that, how would we study the structure of that graph?

Essentially, how is information flow structured?

We'll start with computing transfer entropy. To compute transfer entropy, we actually need to write down an equation. So transfer entropy was originally introduced for discrete time arbitrary order Markov processes.

So suppose we have two Markov processes X and Y.
And we’ll let $X_n$ denote the state of process $X$ at time $n$, $X_{nk}$, where the $k$ is in superscript, to denote the sequence starting from $n - k + 1$ going up to $n$.

So that’s the last $k$ states that the process $X$ visited.

Then, the transfer entropy from $Y$ to $X$, $T_{Y \rightarrow X}$, is the entropy of the future of $X$, conditioned on its past, minus the entropy of the future of $X$ conditioned on its past and the past of the trajectory $Y$.

This means that computing transfer entropy reduces to estimating essentially these entropies.

That means we need to estimate essentially the conditional distributions inside of these parentheses.

That’s easy for certain processes, so for example, if $X$ and $Y$ are Gaussian processes, then really what we’re trying to compute is conditional mutual information, and there are nice equations for conditional mutual information when you have Gaussian random variables.
So if I have three Gaussian random variables, X, Y, Z, possibly multivariate, with joint covariant sigma, then the conditional mutual information between these variables, so the mutual information between X and Y conditioned on Z, is just given by this ratio of log determinants of those covariances. In particular, a common test model used in the transfer entropy literature are linear auto-regressive processes, because a linear auto-regressive process, when perturbed by Gaussian noise, produces a Gaussian process. All of the different joint marginal conditional distributions are all Gaussian, which means that we can compute these covariances analytically, which then means that you can compute the transfer entropy analytically. So these linear auto-regressive processes are nice test cases, 'cause you can do everything analytically. They’re also somewhat disappointing, or somewhat limiting, because in this linear auto-regressive case, transfer entropy is the same as Granger causality. And in this Gaussian case, essentially what we’ve done
is we’ve reduced transfer entropy
to a study of time-lagged correlations.
So this becomes the same as a correlation
based analysis.
We can’t incorporate information beyond the
second moments
if we restrict ourselves to Gaussian processes,
or Gaussian approximations.
The other thing to note is this is strongly
model dependent,
because this particular formula
for computing mutual information
depends on having Gaussian distributions.
In a more general setting, or a more empirical
setting,
you might observe some data.
You don’t know if that data
comes from some particular process,
and you can’t necessarily assume
the conditional distribution is Gaussian.
But we would still like to estimate transfer
entropy,
which leads to the problem of estimating transfer entropy
given an observed time series.
We would like to do this, again, sans model assumptions,
so we don’t want to assume Gaussianity.
This is sort of trivial, again, I stress that,
in discrete state spaces,
because essentially it amounts to counting
occurrences.
But it becomes difficult whenever the state spaces are large
347 00:14:22.920 --> 00:14:25.473 and/or high dimensional, as they often are.
348 00:14:26.340 --> 00:14:28.440 This leads to a couple of different approaches.
349 00:14:28.440 --> 00:14:31.890 So, as a first example, let’s consider spike train data.
350 00:14:31.890 --> 00:14:34.170 So spike train data consists, essentially,
351 00:14:34.170 --> 00:14:38.700 of binning the state of a neuron into either on or off.
352 00:14:38.700 --> 00:14:41.460 So neurons, you can think either in a state zero or one.
353 00:14:41.460 --> 00:14:44.490 And then a pairwise calculation for transfer entropy
354 00:14:44.490 --> 00:14:47.640 only requires estimating a joint probability distribution
355 00:14:47.640 --> 00:14:50.910 over 4 to the k plus l states, where k plus l,
356 00:14:50.910 --> 00:14:53.970 k is the history of x that we remember,
357 00:14:53.970 --> 00:14:55.860 and l is the history of y.
358 00:14:55.860 --> 00:15:00.860 So if the Markov process generating the spike train data
359 00:15:01.350 --> 00:15:04.200 is not of high order, does not have a long memory,
360 00:15:04.200 --> 00:15:06.390 then these k and l can be small,
361 00:15:06.390 --> 00:15:08.160 and this state space is fairly small,
362 00:15:08.160 --> 00:15:09.900 so this falls into that first category,
363 00:15:09.900 --> 00:15:11.520 when we’re looking at a discrete state space
364 00:15:11.520 --> 00:15:13.023 and it’s not too difficult.
365 00:15:14.880 --> 00:15:17.640 In a more general setting, if we don’t try to bin the states
366 00:15:17.640 --> 00:15:19.380 of the neurons to on or off,
367 00:15:19.380 --> 00:15:22.110 for example, maybe we’re looking at a firing rate model,
368 00:15:22.110 --> 00:15:23.970 where we want to look at the firing rates of the neurons,
369 00:15:23.970 --> 00:15:27.210 and that’s a continuous random variable,
370 00:15:27.210 --> 00:15:29.250 then we need some other types of estimators.
So the common estimator used here is a kernel density estimator, or KSG estimator. And this is designed for large, continuous, or high dimensional state spaces, e.g., these firing rate models. Typically, the approach is to employ a Takens delay map, which embeds your high dimensional data in some sort of lower dimensional space, that tries to capture the intrinsic dimension of the attractor that your dynamic process settles onto. And then you try to estimate an unknown density based on this delay map using a k-nearest neighbor kernel density estimate. The advantage of this sort of k-nearest neighbor kernel density is it dynamically adapts the width of the kernel given your sample density. And this has been implemented in some open source toolkits. These are the toolkits that we've been working with. So we've tested this on a couple of different models. And really, I'd say this work, this is still very much work in progress, this is work that Bowen was developing over the summer. And so we developed a couple of different models to test.
The first were these linear auto-regressive networks, and we just used these to test the accuracy of the estimators, because everything here is Gaussian, so you can compute the necessary transfer entropies analytically.

The next, more interesting class of networks are threshold linear networks, or TLNs. These are a firing rate model, where your rate, \( r \), obeys this stochastic differential equation.

So the rate of change in the rate, \( \text{dr}(t) \), is... So you have sort of a leaf term, \(-r(t)\), and then plus, here, this is essentially a coupling, all of this is inside here, the brackets with a plus, this is like a (indistinct) function, so this is just taking the positive part of what’s on the inside. Here, \( b \) is an activation threshold, \( W \) is a wiring matrix, and then \( r \) are those rates again.

And then \( C \) here, that’s essentially covariants for some noise term perturbing this process. We use these TLNs to test the sensitivity of our transfer entropy estimators to common and private noise sources as you change \( C \), as well as how well the transfer entropy network agrees with the wiring matrix.
A particular class of TLNs that were really nice for these experiments are called combinatorial threshold linear networks. These are really pretty new, these were introduced by Carina Curto’s lab this year. And really, this was inspired by a talk I’d seen her give at FACM in May. These are threshold linear networks where the weight matrix here, W, representing the wiring of the neurons, is determined by a directed graph G. So you start with some directed graph G, that’s what’s shown here on the left. This figure is adapted from Carina’s paper, this is a very nice paper if you’d like to take a look at it. And if i and j are not connected, then the weight matrix is assigned one value, and if they are connected, then it’s assigned another value. And the wiring is zero if i equals j. These networks are nice if we want to test structural hypotheses, because it’s very easy to predict from the input graph how the output dynamics of the network should behave. They’re a really beautiful analysis that Carina does in this paper to show that you can produce all these different
interlocking patterns of limit cycles, and multi-step states, and chaos, and all these nice patterns, and you can design them by picking these nice directed graphs. The last class of networks that we’ve built to test are leaky-integrate and fire networks. So here, we’re using a leaky integrate and fire model, where our wiring matrix W is drawn randomly, it’s block-stochastic, which means that it’s Erdos-Renyi between blocks. And it’s a balanced network, so we have excitatory and inhibitory neurons that talk to each other and maintain a balance in the dynamics here. The hope is to pick a large enough scale network that we see properly chaotic dynamics using this leaky integrate and fire model. These tests have yielded fairly mixed results. So the simple tests behave as expected. So the estimators that are used are biased, and the bias typically decays slower than the variance estimation, which means that you do need fairly long trajectories to try to properly estimate the transfer entropy. That said, transfer entropy does correctly identify
causal relationships in simple graphs, and transfer entropy matches the underlying structure used in combinatorial threshold linear networks, so CTLN. Unfortunately, these results did not carry over as cleanly to the leaky integrate and fire models, or to larger models. So what I’m showing you on the right here, this is a matrix where we’ve calculated the pairwise transfer entropy between all neurons in a 150 neuron balanced network. This is shown absolute, this is shown in the log scale. And the main thing I want to highlight, first, taking a look at this matrix, it’s very hard to see exactly what the structure is. You see this banding? That’s because neurons tend to be highly predictive if they fire a lot. So there’s a strong correlation between the transfer entropy between x and y, and just the activity level of x. But it’s hard to distinguish blockwise differences, for example, between inhibitory neurons, excitatory neurons, and that really takes plotting out,
so here, this box and whisker plot on the bottom,
this is showing you if we group entries of this matrix
by type of connection.
So maybe excitatory to excitatory, or inhibitor to excitatory, or so on,
that the distribution of realized transfer entropy
is really different,
but they’re different in sort of subtle ways.
So in this larger scale balanced network,
it’s much less clear whether transfer entropy effectively is equated in some way
with the true connectivity or wiring.
In some ways, this is not a surprise,
because the behavior of the balanced networks
is inherently balanced,
and Erdos-Renyi is inherently in the structure.
But there are ways in which these experiments have revealed
confounding factors that are conceptual factors
that make transfer entropies not an ideal measure,
or maybe not as ideal as it seems,
given the start of this talk.
So for example, suppose two trajectories X and Y
are both strongly driven by a third trajectory Z,
but X responds to Z first.
Well, then the present information about X,
or the present state of $X$, carries information about the future of $Y$, so $X$ is predictive of $Y$.

So $X$ forecasts $Y$, so in the transfer entropy or Wiener causality setting, we would say $X$ causes $Y$, even if $X$ and $Y$ are only both responding to $Z$.

So here, in this example, suppose you have a directed tree where information or dynamics propagate down the tree.

If you look at this node here, $P_j$ and $i$, $P_j$ will react to essentially information traveling down this tree before $i$ does, so $P_j$ would be predictive for $i$, so we would observe an effective connection, where $P_j$ forecasts $i$.

Which means that neurons that are not directly connected may influence each other, and that this transfer entropy, really, you should think of in terms of forecasting, not in terms of being a direct analog to the wiring matrix.

One way around this is to condition on the state of the rest of the network before you start doing some averaging. This leads to some other notions of entropy, so, for example, causation entropy, and this is sort of a promising direction, but it’s not a direction we’ve had time to explore yet.
So that’s the estimation side.

Those are the tools for estimating transfer entropy.

Now, let’s switch gears and talk about that second question I introduced, which is essentially, how do we analyze structure?

Suppose we could calculate a transfer entropy graph. How would we extract structural information from that graph?

And here, I’m going to be introducing some tools that I’ve worked on for a while for describing random structures and graphs. These are tied back to some work I’ve really done as a graduate student, and conversations with Lek-Heng.

So we start in a really simple context, we just have a graph or network. This could be directed or undirected, and we’re gonna require that it does not have self-loops, and that it’s finite.

We’ll let $V$ here be the number of vertices, and $E$ be the number of edges.

Then the object of study that we’ll introduce is something called an edge flow. An edge flow is essentially a function on the edges of the graph.
and returns a real number. And this is an alternating function, so if I take $f(i, j)$, that’s negative $f(j, i)$, because you can think of $f(i, j)$ as being some flow, like a flow of material between i and j, hence this name, edge flow. This is analogous to a vector field, because this is analogous to the structure of a vector field on the graph, and represents some sort of flow between nodes. Edge flows are really sort of generic things. So you can take this idea of an edge flow and apply it in a lot of different areas, because really all you need is you just need the structure of some alternating function on the edges of a graph. So I’ve read papers, and worked in a bunch of these different areas. Particularly, I’ve focused on applications of this in game theory, in pairwise and social choice settings, in biology and Markov chains. And a lot of this project has been attempting to take this experience working with edge flows in, for example, say, non-equilibrium thermodynamics, or looking at pairwise preference data,
and looking at a different application area here to neuroscience. Really, you can think about the edge flow, or relevant edge flow in neuroscience, you might be asking about asymmetries in wiring patterns, or differences in directed influence or causality, or, really, you can think about these transfer entropy quantities. This is why I was excited about transfer entropy. Transfer entropy is inherently directed notion of information flow, so it’s natural to think that if you can calculate things like the transfer entropy, then really, what you’re studying is some sort of edge flow. Edge flows often are subject to the same common questions. So if I want to analyze the structure of an edge flow, there’s some really big global questions that I would often ask, that get asked in all these different application areas. One common question is, well, does the flow originate somewhere and end somewhere? Are there sources and sinks in the graph? Another is, does it circulate? And if it does circulate, on what scales, and where?
If you have a network that’s connected to a whole exterior network, for example, if you’re looking at some small subsystem that’s embedded in a much larger system, as is almost always the case in neuroscience, then you also need to think about what passes through the network. So, is there a flow or current that moves through the boundary of the network, and is there information that flows through the network you’re studying? And in particular, if we have these different types of flow, if flow can originate in source and end in sinks, if it can circulate, if it can pass through, can we decompose the flow into pieces that do each of these?

Those questions lead to a decomposition. So here, we’re going to start with a simple idea.

We’re going to decompose an edge flow by projecting it onto orthogonal subspaces associated with some graph operators. Generically, if we consider two linear operators, where the product A times B equals zero, then the range of B must be contained in the null space of A, which means that I can express essentially any set of real numbers,
so you can think of this as being the vector space of possible edge flows, as a direct sum of the range of B, the range of A transpose, and the intersection of the null space of B transpose and the null space of A.

This blue subspace, this is called the harmonic space, and this is trivial in many applications if you choose A and B correctly.

So there’s often settings where you can pick A and B so that these two null spaces have no intersection, and then this decomposition boils down to just separating a vector space into the range of B and the range of A transpose.

In the graph setting, our goal is essentially to pick these operators to be meaningful things, that is, to pick graph operators so that these subspaces carry a meaningful, or carry meaning in the structural context.

So let’s think a little bit about graph operators here.

So, let’s look at two different classes of operators.

So we can consider matrices that have E rows and n columns, or matrices that have l rows and E columns, where again, E is the number of edges in this graph.
If I have a matrix with E rows, then each column with a matrix has as many entries as there are edges in the graph, so it can be thought of as itself an edge flow. So you can think that this matrix is composed of a set of columns, where each column is some particular motivic flow.

In contrast, if I look at a matrix where I have E columns, then each row of the matrix is a flow motif, so products against M evaluate inner products against specific flow motifs. That means in this context, if I look at the range of this matrix, this is really a linear combination of a specific subset of flow motifs, and in this context, if I look at the null space of the matrix, I'm looking at all edge flows orthogonal to that set of flow motifs. So here, if I look at the range of a matrix with E rows, that subspace is essentially modeling behavior similar to the motifs, so if I pick a set of motifs that flow out of a node, or flow into a node, then this range is going to be a subspace of edge flows that tend to originate in sources and end in sinks. In contrast, here, the null space of M,
that’s all edge flows orthogonal to the flow motifs, so it models behavior distinct from the motifs. Essentially, this space asks what doesn’t the flow do, whereas this space asks what does the flow do. Here is a simple, very classical example. And this goes all the way back to, you can think, like Kirchhoff electric circuit theory.

We can define two operators. Here, G, this is essentially a gradient operator. And if you’ve taken some graph theory, you might know this as the edge incidence matrix. This is the matrix which essentially records the endpoints of an edge, and evaluates differences across it. So for example, if I look at this first row of G, this corresponds to edge I in the graph, and if I had a function defined on the nodes in the graph, products with G would evaluate differences across this edge. If you look at its columns, each column here is a flow motif. So for example, this highlighted second column, this is entries 1, -1, 0, -1, if you carry those back to the edges, that corresponds to this specific flow motif. So here, this gradient, it’s adjoint,
so essentially a divergence operator, which means that the flow motifs are unit in flows
or unit out flows for specific nodes, like what’s shown here.
You can also introduce something like a curl operator. The curl operator evaluates path sums around loops.
So this row here, for example, this is a flow motif corresponding to the loop labeled A in this graph.
You could certainly imagine other operators build other motifs.
These operators are particularly nice, because they define principled subspaces.
So if we apply that generic decomposition, then we could say that the space of possible edge flows, RE,
can be decomposed into the range of the gradient operator.
the range of the curl transpose, and the intersection of their null spaces, into this harmonic space.
This is nice, because the range of the gradient, that’s flows that start and end somewhere, those are flows that are associated
with motion (indistinct) potential.
So these, if you’re thinking physics, you might say that these are conservative.
these are flows generated by voltage
if you’re looking at an electric circuit.
These cyclic flows, while these are the flows and range of the curl transpose, and then this harmonic space, those are flows that enter and leave the network without either starting or ending at a sink or a source, or circulating. So you can think that really, this decomposes the space of edge flows into flows that start and end somewhere inside the network, flows that circulate within the network, and flows that do neither, i.e. flows that enter and leave the network. So this accomplishes that initial decomposition I'd set out at the start. Once we have this decomposition, then we can evaluate the sizes of the components of the decomposition to measure how much of the flow starts and ends somewhere, how much circulates, and so on. So, we can introduce these generic measures, where given some operator M, we decompose the space of edge flows into the range of M and the null space of M transpose, which means we can project f onto these subspaces, and then evaluate the sizes of these components.
and that’s a way of measuring how much of the flow behaves like the flow motifs contained in this operator, and how much doesn’t. So, yeah. So that lets us answer this question, and this is the tool that we’re going to be using as our measurable. Now, that’s totally easy to do, if you’re given a fixed edge flow and a fixed graph. If you have a fixed graph, you can build your operators, you choose the motifs, you have fixed edge flow, you just project the edge flow onto the subspaces, span by those operators, and you’re done. However, there are many cases where it’s worth thinking about a distribution of edge flows, and then expected structures given that distribution. So here, we’re going to be considering random edge flows, for example, an edge flow of capital F. Here, I’m using capital letters to denote random quantities sampled from an edge flow distribution. So this is the distribution of possible edge flows. And this is worth thinking about
because many generative models are stochastic.

They may involve some random seed, or they may, for example, like that neural model, or a lot of these sort of neural models, be chaotic.

so even if they are deterministic generative models, the output data behaves as though it’s been sampled from a distribution.

On the empirical side, for example, when we’re estimating transfer entropy, or estimating some information flow, then there’s always some degree of measurement error.

which really means that from a Bayesian perspective, we should be thinking that our estimator is a point estimate drawn from some posterior distribution of edge flows, and that we’re back in the setting where, again, we need to talk about a distribution.

Lastly, this random edge flow setting is also really important if we want to compare the null hypotheses.

Because often, if you want to compare to some sort of null hypothesis, it’s helpful to have an ensemble of edge flows to compare against,
which means that we would like to be able to
talk about
expected structure under varying distribu-
tional assumptions.
If we can talk meaningfully about random
edge flows,
then really what we can start doing
is we can start bridging the expected structure
back to the distribution.
So what we’re looking for
is a way of explaining generic expectations
of what the structure will look like
as we vary this distribution of edge flows.
You could think that a particular dynamical
system generates a wiring pattern, that generates
firing dynamics,
those firing dynamics determine
some sort of information flow graph,
and then that information flow graph
is really a sample from that generative model,
and we would like to be able to talk about
what would we expect
if we knew the distribution of edge flows
about the global structure.
That is, we’d like to bridge global structure
back to this distribution.
And then, ideally, you’d bridge that distribu-
tion
back to the generative mechanism.
And this is a project for future work.
Obviously, this is fairly ambitious.
However, this first point is something you can do really in fairly explicit detail, and that’s what I would like to spell out with the end of this talk, is how do you bridge global structure back to a distribution of edge flows. So yeah.

So that’s our main question. How does the choice of distribution influence the expected global flow structure? So first, let’s start with a lemma. Suppose that we have a distribution of edge flows with some expectation $f_{\bar{\bar{b}}}$, and some covariance, here I’m using double bar $V$ to denote covariance.

We’ll let $S$ contained in the set of... $S$ will be a subspace contained within the vector space of edge flows, and we’ll let $P_S$ be the orthogonal projector onto $S$. Then $F_S$, that’s the projection of $F$ onto this subspace $S$,

the expectation of its norm squared is the norm of the expected flow projected onto $S$ squared, so this is essentially the expectation of the sample is the measure evaluated with the expected sample. And then plus a term that involves an inner product.
between the projector on the subspace and the covariance matrix for the edge flows. Here, this denotes the matrix inner product, so is just the sum over all \(ij\) entries. What’s nice about this formula is, at least in terms of expectation, it reduces this study of the bridge between distribution and network structure to a study of moments, right? Because we’ve replaced a distributional problem here with a linear algebra problem that’s posed in terms of this projector, the projector out of the subspace \(S\), which is determined by the topology of the network.

And the variance in that edge flow, which is determined by your generative model. Well, you might say, "Okay, well, fine, this is a matrix inner product, we can just stop here."

We could compute this projector. We could sample a whole bunch of edge flows to compute this covariance. So you can do this matrix inner product.” But I’m sort of greedy, because I suspect that you can really do more with this inner product.

So I’d like to highlight some challenges associated with this inner product. So first, let’s say I asked you to design a distribution
So for example, I said I want you to pick a generative model, or design a distribution of edge flows, that when I sample edge flows from it, their expected structures match some expectation. It’s not obvious how to do that given this formula. It’s not obvious in particular, because these projectors, like the projector onto subspace S, typically depend in fairly non-trivial ways on the graph topology. So small changes in the graph topology can completely change this projector. In essence, it’s hard to isolate topology from distribution.

You could think that this inner product, if I think about it in terms of the ij entries, while easy to compute, is not easy to interpret, because i and j are somewhat arbitrary indexing. And obviously, really, the topology of the graph, it’s not encoded in the indexing, it’s encoded in the structure of these matrices. So in some ways, what we really need is a better basis for computing this inner product. In addition, computing this inner product just may not be empirically feasible, because it might not be feasible.
to estimate all these covariances. There’s lots of settings where, if you have a random edge flow, it becomes very expensive to try to estimate all the covariances in this graph, or sorry, in this matrix, because this matrix has as many entries as there are pairs of edges in the graph. And typically, that number of edges grows fairly quickly in the number of nodes in the graph. So in the worst case, the size of these matrices goes not to the square of the number of nodes in the graph, but the number of nodes in the graph to the fourth, so this becomes very expensive very fast. Again, we could try to address this problem if we had a better basis for performing this inner product, because we might hope to be able to truncate somewhere in that basis, and use a lower dimensional representation. So, to build there, I’m going to show you a particular family of covariances. We’re going to start with a very simple generative model. So let’s suppose that each node of the graph is assigned some set of attributes, a random vector X, sampled from a... So you can think of trait space, a space of possible attributes.
And these are sampled i.i.d. In addition, we'll assume there exists an alternating function \( f \), which accepts pairs of attributes, and returns a real number. So this is something that I can evaluate on the endpoints of an edge, and return an edge flow value. In this setting, everything that I've shown you before simplifies. So if my edge flow \( F \) is drawn by first sampling a set of attributes, and then plugging those attributes into functions on the edges, then the mean edge flow is zero, so that \( f \) bar goes away, and the covariance reduces to this form. So you get a standard form, where the covariance and the edge flow are functions of two scalar quantities, that's \( \sigma^2 \) and \( \rho \), these are both statistics associated with this function and the distribution of traits. And then some matrices, so we have an identity matrix, and we have this gradient matrix showing up again. This is really nice, because when you plug it back in, we try to compute, say, the expected sizes of the components,
this matrix inner product that I was complaining about before, this whole matrix inner product simplifies. So when you have a variance that’s in this nice, simple, canonical form, then the expected overall size of the edge flow, that’s just sigma squared, the expected size projected onto that conservative subspace, that breaks into this combination of the sigma squared and the rho, those are some simple statistics. And then V, E, L, and E, those are just essentially dimension counting on the network. So this is the number of vertices, the number of edges, and the number of loops, the number of loops, that’s the number of edges minus the number of vertices plus one. And similarly, the expected cyclic size, or size of the cyclic component, reduces to, again, this scalar factor in terms of the statistics, and some dimension counting topology related quantities. So this is very nice, because this allows us to really separate the role of topology from the role of the generative model. The generative model determines sigma and rho, and topology determines these dimensions. It turns out that the same thing is true.
even if you don’t sample the edge flow using this trait approach, but the graph is complete.

So if your graph is complete, then no matter how you sample your edge flow, for any edge flow distribution, exactly the same formulas hold, you just replace those simple statistics with estimators for those statistics given your sampled flow. And this is sort of a striking result, because this says that this conclusion that was linked to some specific generative model that was linked to some specific generative model with some very specific assumptions, right, we assumed it was i.i.d., extends to all complete graphs, regardless of the actual distribution that we sampled from.

Up until this point, this is kind of just an algebra miracle. And one of the things I’d like to do at the end of this talk is explain why this is true, and show how to generalize these results. So to build there, let’s emphasize some of the advantages of this. So first, the advantages of the model, it’s mechanically plausible in certain settings.
it cleanly separated the role of topology and distribution, and these coefficients that had to do with topology, these are just dimensions, these are non negative quantities, so it’s easy to work out monotonic relationships between expected structure and simple statistics, of the edge flow distribution. The fact that you can do that enables more general analysis. This was an experiment where we trained a set of agents to play a game using a genetic algorithm, and then we looked at the expected sizes of cyclic and acyclic components in a tournament among those agents. And you can actually predict these curves because it was possible to predict the dynamics of these simple statistics, this sigma and this rho. So this is a really powerful analytical tool, but it is limited to this particular model. In particular, it only models unstructured cycles, so if you look at the cyclic component generated by this model,
it just looks like random noise that’s been projected onto the range of the current transpose. It’s limited to correlations on adjacent edges, so we only generate correlations on edges that share an endpoint, because you could think all of the original random information comes from the endpoints. And then, in some ways, it’s not general enough. We can’t parameterize all possible expected structures by picking sigma and rho. And we lack some notion of sufficiency, i.e. if the graph is not complete, then this nice algebraic property fails to hold. So if the graph is not complete, projection onto the family of covariances parameterized in this fashion changes the expected global structure. So we would like to address these limitations. And so our goal for the next part of this talk is to really generalize these results. To generalize, we’re going to switch our perspective a little bit. So I’ll recall this formula, that if we generate our edge flow
by sampling quantities on the endpoints, and then plugging them into functions on the edges, then you necessarily get a covariance that’s in this two parameter family, where I have two scalar quantities associated with the statistics of the edge flow, that’s this sigma and this rho, and I have some matrices that are associated with the topology of the network in the subspaces I’m projecting onto. These are related to a different way of looking at the graph. So I can start with my original graph, and then I can convert it to an edge graph, where I have one node per edge in the graph, and nodes are connected if they share an endpoint. You can then assign essentially signs to these edges based on whether the edge direction chosen in the original graph is consistent or inconsistent at the node that links two edges. So for example, edges 1 and 2 both point in to this node, so there’s an edge linking 1 and 2 in the edge graph with a positive sign. This essentially tells you that the influence of random information assigned on this node linking 1 and 2.
would positively correlate the sample edge flow on edges 1 and 2.

Then, this form, what this form for covariance matrices says is that we're looking at families of edge flows that have correlations on edges sharing an endpoint, so edges at distance one in this edge graph, and non-adjacent edges are entirely independent of each other.

Okay. So that's essentially what the trait-performance model is doing, is it's parameterizing a family of covariance matrices, where we're modeling correlations at distance one, but not further in the edge graph.

So then the natural thought for how to generalize these results is to ask, "Can we model longer distance correlations to this graph?"

To do so, let's think a little bit about this matrix that's showing up inside the covariances, so we have a gradient times a gradient transpose. This is in effect a Laplacian for that edge graph.

And you can do this for other motifs.

If you think about different motif constructions,
1123 00:46:34.710 --> 00:46:38.400 essentially if you take a product of M transpose times M,
1124 00:46:38.400 --> 00:46:41.070 that will generate something that looks like a Laplacian
1125 00:46:41.070 --> 00:46:44.070 or an adjacency matrix for a graph
1126 00:46:44.070 --> 00:46:47.250 where I’m assigning nodes to be motifs,
1127 00:46:47.250 --> 00:46:50.190 and looking at the overlap of motifs.
1128 00:46:50.190 --> 00:46:54.840 And if I look at M times M transpose,
1129 00:46:54.840 --> 00:46:58.650 and I’m looking at the overlap of edges via shared motifs.
1130 00:46:58.650 --> 00:46:55.300 So these operators you can think about as being Laplacians
1131 00:46:55.300 --> 00:46:58.650 for some sort of graph
1132 00:46:58.650 --> 00:47:01.413 that’s generated from the original graph motifs.
1133 00:47:01.413 --> 00:47:03.630 Like any adjacency matrix,
1134 00:47:03.630 --> 00:47:11.040 powers of something like G, G transpose minus 2I,
1135 00:47:11.040 --> 00:47:13.800 that would model connections along longer paths,
1136 00:47:13.800 --> 00:47:15.810 along longer distances in these graphs
1137 00:47:15.810 --> 00:47:18.710 associated with motifs, in this case, with the edge graph.
1138 00:47:18.710 --> 00:47:21.240 So our thought is, maybe, well,
1139 00:47:21.240 --> 00:47:22.890 we could extend this trait performance
1140 00:47:22.890 --> 00:47:24.630 family of covariance matrices
1141 00:47:24.630 --> 00:47:26.610 by instead of only looking at
1142 00:47:26.610 --> 00:47:30.750 a linear combination of an identity matrix and this matrix,
1143 00:47:30.750 --> 00:47:32.190 we could look at a power series.
1144 00:47:32.190 --> 00:47:36.600 So we could consider combining powers of this matrix.
1145 00:47:36.600 --> 00:47:39.390 And this would generate this family of matrices
that are parameterized by some set of coefficients (indistinct)... Dr. Strang.

I apologize, I just wanted to remind you that we have a rather tight time limit.

So here, the idea is to parameterize this family of matrices by introducing a set of polynomials with coefficients alpha, and then plugging into the polynomial that’s generated by the Laplacian that’s generated by the graph motifs. The adjacency matrix generated by the graph motifs.

we’re interested in. And that trait performance result, that was really just looking at the first order case here,

that was looking at a linear polynomial with these chosen coefficients.

This power series model is really nice analytically.

So if we start with some graph operator M, and we consider the family of covariance matrices.

generated by plugging M, M transpose into some polynomial and power series, then this family of matrices is contained within the span of powers of M, M transpose.

You can talk about this family in terms of combinatorics.
So, for example, if we use that gradient times gradient transpose minus twice the identity, then powers of this is essentially, again, path counting, so this is counting paths of length \( n \). You can also look at things like the trace of these powers. If you look at the trace series, that’s the sequence where you look at the trace of powers of these, essentially, these adjacency matrices. This is doing some sort of loop count, where we’re counting loops of different length. And you can think of this trace series, in some sense, as controlling amplification of self-correlations within the sampled edge flow. Depending on the generative model, we might want to use different operators for generating these families. So for example, going back to that synaptic plasticity model with coupled oscillators, in this case, using the gradient to generate the family of covariance matrices is not really the right structure, because the dynamics of the model have these natural cyclic connections. So it’s better to build the power series using the curl.
So depending on your model, you can adapt this power series family by plugging in a different graph operator. Let’s see now what happens if we try to compute the expected sizes of some components using a power series of this form. So, if the variance, or covariance matrix for edge flow is a power series in, for example, the gradient, gradient transpose, then the expected sizes of the measures can all be expressed as linear combinations of this trace series and the coefficients of the original polynomial. For example, the expected cyclic size of the flow is just the polynomial evaluated at negative two, multiplied by the number of loops in the graph. And this, this really generalizes that trait performance result, because the trait performance result is given by restricting these polynomials to be linear. This, you can extend to other bases. But really, what this accomplishes is by generalizing trait performance, we achieve this generic properties that it failed to have. So in particular, if I have an edge flow subspace $S$
spanned by the flow motifs stored in some operator $M$,
then this power series family of covariances associated with the Laplacian, that is, $M$ times $M$ transpose,
is both expressive, in the sense that
for any non negative $a$ and $b$,
I can pick some alpha and beta
so that the expected size of the projection of $F$ onto the subspaces $a$, and the projected size of $F$ on the subspace orthogonal to $S$ is $b$ for any covariance in this power series family.
And it’s sufficient in the sense that
for any edge flow distribution with mean zero,
if $C$ is the matrix nearest to $V$ in Frobenius norm,
then these inner products computed in terms of $C$
are exactly the same as inner products
computed in terms of $V$, so they directly predict the structure,
which means that if I use this power series family,
discrepancies off of this family
don’t change the expected structure.
Okay.
So, I know I’m short on time here,
so I’d like to skip, then, just to the end of this talk.
There’s further things you can do with this, this is sort of really nice mathematically. You can build an approximation theory out of this, and study it for different random graph families, how many terms in these power series you need. And those terms define some nicer simple minimal set of statistics, to try to estimate structure. But I’d like to really just get to the end here, and emphasize the takeaways from this talk. So the first half of this talk was focused on information flow. What we saw is that information flow is a non-trivial, but well studied estimation problem. And this is something that, at least on my side, is a work in progress with students. Here, the, in some ways, the conclusion of that first half would be that causation entropy may be a more appropriate measure than TE when trying to build these flow graphs to apply these structural measures to. Then, on the structural side, we can say that power series families, this is a nice family of covariance matrices. It has nice properties that are useful empirically,
because they let us build global correlation structures from a sequence of local correlations from that power series. If you plug this back into the expected measures, you can recover monotonic relations, like in that limited trait performance case. And truncation of these power series reduces the number of quantities that you would actually need to measure. Actually, to a number of quantities that can be quite small relative to the graph, and that’s where this approximation theory comes in. One way, maybe to sort of summarize this entire approach, is what we’ve done is by looking at these power series built in terms of the graph operators, it provides a way to study inherently heterogeneous connections, or covariances, or edge flows distributions, using a homogeneous correlation model that’s built at multiple scales by starting with local scale and then looking at powers. In some ways, this is a common... I ended a previous version of this talk with, I still think that this structural analysis is, in some ways, a hammer seeking a nail,
and that this information flow construction, this is work in progress to try and build that nail.

So thank you all for your attention. I’ll turn it now over to questions.

Thank you so much for your talk. For those of you on Zoom, you’re welcome to keep up the conversations, but unfortunately we have to clear the room, so I do apologize. Dr. Strang? Am I muted?

Dr. Strang?

Oh, yes, yeah.

Okay, do you mind if people...

We have to clear the room, do you mind if people email you if they have questions?

I’m sorry, I couldn’t hear the end of the question.

Do I mind if...

We have to clear the room,

do you mind if people email you if they have questions,

No, no, not at all.

So I do apologize, they are literally
1319 00:54:54.466 --> 00:54:56.760 (indistinct) the room right now.
1320 00:54:56.760 --> 00:54:59.100 <v ->Okay, no, yeah, that’s totally fine.</v>
1321 00:54:59.100 --> 00:55:00.660 <v Instructor>Thank you.</v>
1322 00:55:00.660 --> 00:55:02.820 And thanks again for a wonderful talk.
1323 00:55:02.820 --> 00:55:03.653 <v ->Thank you.</v>