CAUSAL MAP GARDEN

Task 3 – Answering questions – General

Contents

🌻 Causal mapping produces models you can query to answer questions

Chapter contents.

20 Sep 2025

> The fundamental output of causal mapping is a database of causal links. If there are not too many links, this database can be visualised "as-is" in the form of a causal map or network. But usually there are too many links for this to be very useful, so we apply filters.

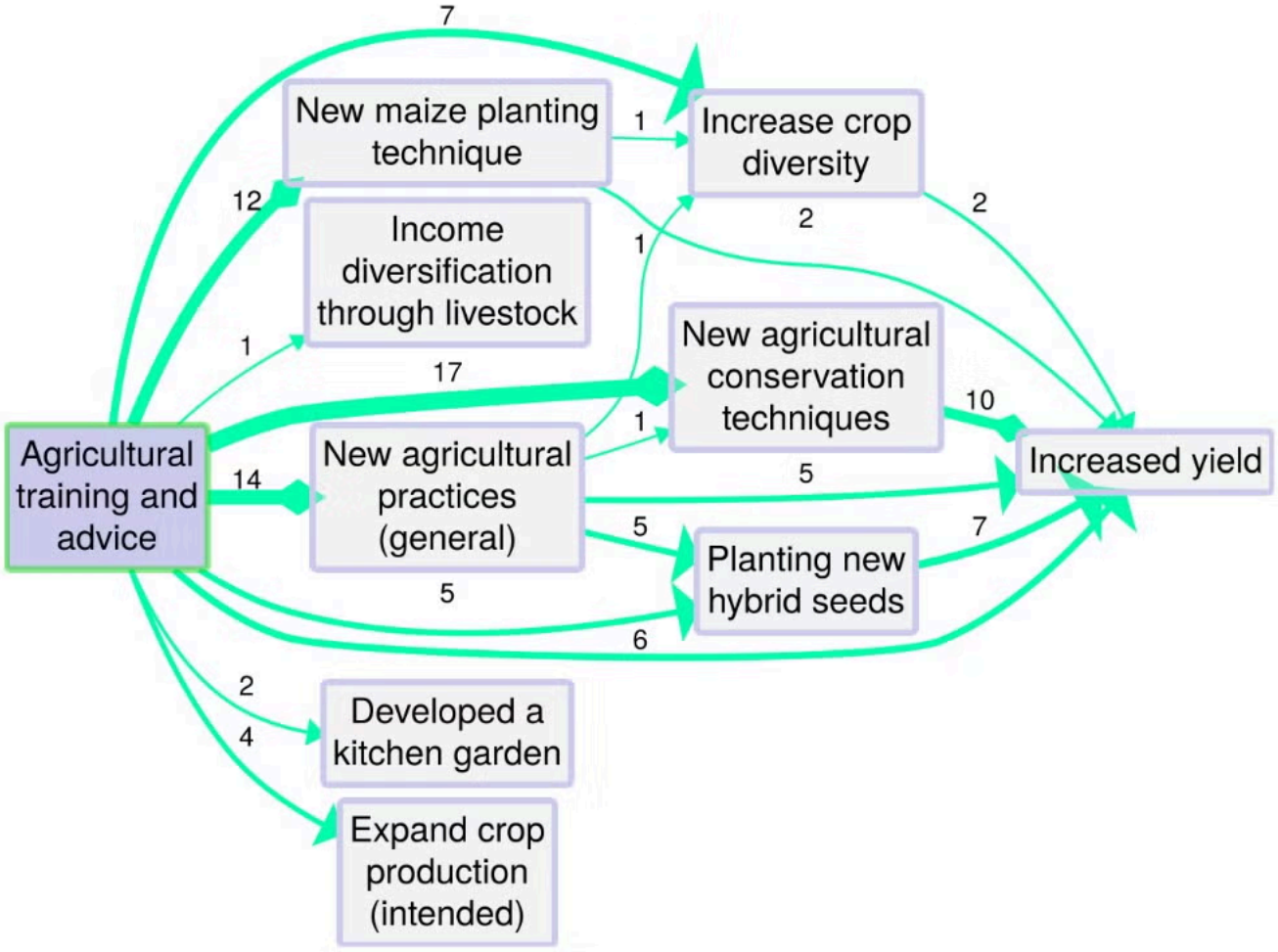
By applying filters and other algorithms, a causal map can be queried in different ways to answer different questions, for example to simplify it, to trace specific causal paths, to identify significantly different sub-maps for different groups of sources, etc.

As explained on the Causal Mapping website:

"*A global causal map resulting from a research project can contain a large number of links and causal factors. By applying filters and other algorithms, a causal map can be queried in different ways to answer different questions, for example to simplify it, to trace specific causal paths, to identify significantly different sub-maps for different groups of sources, etc.*"

The figure below shows a map from the application Causal Map, showing coded causal statements for a project that provided farmers with agricultural training and advice in order to increase crop yields. The

map has been filtered to show only outcomes downstream of the influence factor ‘Agricultural training and advice’. Numbers shown indicate how many times the links were mentioned across all interviews.



Source: BDSR, 2021, p 4

Here is the same kind of question inside the Causal Map app, using the shared `example-original` project. This bookmark asks a downstream question: *what follows from Increased Knowledge?*

Consequences of Increased Knowledge in example-original (bookmark #262)

Bookmark #262 — looking downstream from one factor in a larger coded project. Open in app

PAGES IN THIS CHAPTER

Outputs of QDA

The most controversial feature of causal maps is transitivity

Ways to causal inference

Epistemic logic does not help us with reasoning about causal maps

We can reason about causal maps using a logic of evidence

Causal maps are knowledge graphs, but with wings

Quality assurance at each step of the causal coding workflow

Assessing quality or robustness of evidence for a causal link based on a bundle of coterminal causal claims

The product of (causal) qualitative coding can be a model you can query

The transitivity trap

Outputs of QDA

2 Oct 2025

The logic of QDA: you've done your analysis, now what?

The result of qualitative data analysis can be understood as, at least, some kind of qualitative theory or model at least of the sources' beliefs, with at least some possibility of generalising beyond them. But it can be hard to know what to do with the results of an emergent qualitative text analysis. There is no clear decision procedure: we can ask the author, and the answer is: some explanation, i.e. more text. In more reproducible approaches we do get some more structured outputs such as tables of frequencies. Some authors such as Mayring see these kinds of outputs as an important analysis result. QDA software is often used to capture and structure and even make inferences with these kinds of outputs.

In the logic of (non-causal) reproducible QDA, we can do things like this:

- count occurrences of concepts, and use ordinary arithmetic to report eg which of two concepts was more common
- count co-occurrences of two concepts, and construct measures like association between concepts, and more generally combine and query occurrences with boolean logic
- create case/code matrices
- report relationships between sources and concepts, for example to compare codings of one concept for different genders
- reason about concepts, for example to deduce that an occurrence of "lion" is also an occurrence of "mammal", either relying on our implicit understanding of the concepts or through the explicit declaration of a parent-child relationship.

Of course frequency statistics are notoriously unstable, because they depend on our decisions about granularity and chunking. If I have a codebook which has 100 different codes for cats and only 1 code for dogs, we may conclude that dogs were mentioned in the text more often than any other animal-concept even if cat-concepts were mentioned more often in combination. This is one reason why reasoning with these kinds of outputs can never be merely automated. There always has to be a "human in the loop".

Nevertheless the point is that we can understand the output of QDA coding as some proportion of "more text", which itself needs to be interpreted by humans, and a complementary proportion of machine-readable, structured output which can be used to ask and answer questions (Which are the overarching themes? How much does climate anxiety come up as a theme? Who mentions it most?) at least somewhat independently of human guidance.

QDA logic can also be extended beyond the simple logic of frequencies and occurrences to apply (special kinds of) codes which have additional explicit rules associated with them, such as code weighting (as for example in MaxQDA). This means we can for example apply codes like “somewhat happy” or ‘very unhappy’ which enable us to say that the expression of happiness in one case is stronger than the other, or (if we also allow coding for time) that happiness increases or decreases over time. These extra deductions we can make come free with the (implicit or explicit) underlying ordinal logic of comparison of intensity.

QDA without coding

Coding does not have to be central to qualitative data analysis (Morgan 2025; Nguyen-Trung & Nguyen 2025). ...

Related

- chapter intro

References

Morgan (2025). *Query-Based Analysis: A Strategy for Analyzing Qualitative Data Using ChatGPT*. https://doi.org/10.1177/10497323251321712.

Nguyen-Trung, & Nguyen (2025). *Narrative-Integrated Thematic Analysis (NITA): AI-Supported Theme Generation Without Coding*. https://doi.org/10.31219/osf.io/7zs9c_v1.

🌻 The most controversial feature of causal maps is transitivity

9 Oct 2025

How does causal inference work in a causal *network*?

When is a pathway not just a link?

The logic around how links might combine into pathways and what that means for evaluation, that's the most exciting part. e.g. how might this intervention influence an outcome which might be multiple steps downstream of it?

From

> a —>b

and

> b —> c

what can we conclude about

> a —> c. ?

For example, if the relation —> means "causes", when and under what circumstances can we conclude that a causes c?

Once we know the inference rules for a network, in particular the transitivity rule, we can infer all kinds of useful things about it.

There is a whole library of thinking about causal reasoning within a statistical or probablistic network.

There is less written about qualitative causality within a qualitative causal network.

But our problem is harder again: to reason with what we call a causal map, where the links are about **beliefs about** or **evidence for** a causal connection.

We can reason about causal maps using a logic of evidence

Related

- chapter intro

Ways to causal inference

9 Oct 2025

There are different ways to set up a table like this. Some of these are more like methods, some are more like frameworks. Some overlap. Most do not exist only or even primarily to conduct causal inference.

Commonalities

Almost all these meta-models presuppose that causal knowledge can be captured in chunks, and these chunks can be joined together in chains / networks in order to consider indirect effects, ie if X causes Y and Y causes Z, what can we say about the causal effect of X on Z? Everything doesn't depend on *everything* else: in a large causal network, to make causal inferences about Z the only causal factors we need to consider are those for which there is a causal chain leading to Z, and the only direct effects on Z are the factors which have arrows directly leading to Z.

"Portability" means that the knowledge that Xs can influence Ys can be applied in multiple contexts, not just on one particular occasion.

Traditionally, more qualitative approaches like QCA are less keen on portability.

Some of the models eg SEM are particularly interested in calculating indirect effects in chains and networks. In others eg QCA the concept of a causal chain is less clear, and there is little portability / generalisibility.

Table

	Notes	Prerequisites	Inference	Metaphor	
Ordinary reasoning		I already know that Xs are likely to cause Ys, and this X happened and Y followed it in just the way you'd expect. I don't exactly have a ready-made theory about X's and Y's but I have similar	It is quite likely that X causally influenced Y.	Both the foundation on which all other methods stand and the cement which holds them together	

		theories about things similar to Xs and Ys and some other knowledge which suggest that the these similar			
Ordinary reasoning: Modus Operandi		The signatures of all the other possible explanations of Y were absent and the signature of X was present. A signature is evidence of a chain from X to Y or other evidence that X was active. Relies on pre-existing causal knowledge Argues also that counterfactuals may be irrelevant; the patient might also have died if they hadn't had the heart attack, but the heart attack was the direct physical cause.			Scriven
Ordinary reasoning: Multiple lines and levels of evidence (MLLE)					
Ordinary reasoning: Causal mapping		Not really a method of causal inference at all, but a way to organise medium or large datasets of pre-existing (possibly conflicting or			

		overlapping) causal claims or inferences made by a set of sources. Says nothing about portability because it relies on its sources to decide on what			
Successionist		- Many observations that Ys follow Xs	May in practice be used to support but not warrant the conclusion that the Xs caused the Ys. Might be used in evaluation as a hoop test.		Hume
Classical Statistical		Not possible			
Regression discontinuity					
Counterfactual		Can be used for **disproving** a causal inference: if sometimes Xs appear but Ys do not appear, this X cannot be the cause of this Y.	X caused Y		
DAGs / SEM		Level 3: A model in which the effect of X on Y is significant Substantive theory			Pearl
Configurational; QCA, fQCA		A dataset showing the co-occurrence and non-occurrence of X1, X2, X3 ... and Y	This combination of Xs may have caused Y in this case.	Causal package	Ragin

		A case in which X1, X2, X3 etc and Y occur/don't occur in a pattern corresponding to the dataset Substantive considerations backing up the dataset Possible consolidation of the information into more parsimonious structures. Usually not used to establish general causal laws and then make specific causal inferences but only to make causal inferences within a dataset. The inference that X1, X2, X3 actually *cause* Y rather than just being associated with it is not based only on the configuration but has to be provided by other substantive knowledge, so perhaps QCA is not strictly a method of causal inference.			
INUS		Special case of the above in which X1, X2, X3 … are necessary parts of a package P which is	X1 was an INUS-cause of Y (and so was X2 etc)	Causal package	

		sufficient for Y, i.e. (X1 & X2 & X3...) is sufficient for Y.			
Realist		Substantive knowledge that a mechanism M in a context C, when triggered (X) causes Y. M and C were present, X was activated, Y happened.	X was a cause of Y	A machine or mechanism	
Process oriented; Process tracing					
Physical causality		“people’s “operative reasons” for doing what they do are the physical actions.” “A design whose purpose is to determine impact will be considered qualitative if it relies on something other than evidence for the counterfactual to make a causal inference. It is qualitative first in the positive sense that it rests on demonstrating a quality-in the present treatment the quality will be a physical connection in the natural world between the proposed cause		Touching	Mohr

		and effect-and second in the negative sense that it is not quantitative, that is, it does not rely on a treatment variable, or on comparing what is with an estimate of what would otherwise have been.”			
Dispositional					
Radical systems eg Quinn Patton?		Not possible			

Related

- chapter intro

🌻 Epistemic logic does not help us with reasoning about causal maps

9 Oct 2025

(An example of kind-of qualitative causal logic, with a focus on groups: Castellani et al. (2025)])

> From (Powell et al. 2024)

Seen as models of the world, causal maps, like systems maps, are fallible but useful: We can use inference rules (which are explicitly set out in FCMs, SDs, BBNs and CLDs and are implicit in other related approaches), and in particular, transitivity rules, to make deductions about the world.

There are at least three problems of transitivity which we need to think about

1. Given that A influences B and B influences C, does A influence C?
2. Given that P believes that A influences B and P believes that B influences C, does P believe that A influence C?
3. Given that someone believes that A influences B and someone else P believes that B influences C, does someone (who? we? the people?) believe that A influence C?

So if A causes B and B causes C, causal logic might tell us the answer to 1) under what circumstances A causes C.

Seen as models of individuals' causal beliefs, we can arguably use analogous rules to make deductions about what individuals believe, or ought to believe, given what else they believe.

There is a thing called epistemic logic which is a strange shadow of causal logic. Can it help us answer 2 and 3?

But epistemic logic is a strange thing.

If a person P believes that A causes B and B causes C, epistemic logic tells us what P believes about A causing C *if they were a rational person*. Whereas, facts about what people actually do believe is a branch of psychology.

In the last decades, thinkers like Daniel Kahneman have shown that in this sense, humans are so far from rational that it does not make sense even to start off with a rationality assumption and then add some corrections.

It would be great to use causal maps to infer, given a bunch of information about different people's causal beliefs, what they believe about *other* causal connections. That would be really useful. But it is hard.

There is a much easier way to reason with causal maps which is also vital for evaluators: to reason about **evidence**.

We can reason about causal maps using a logic of evidence

Related

- chapter intro

References

Castellani, Schimpf, Wistow, Caden, Agarwal, & Barbrook-Johnson (2025). *Case-Based Systems Mapping: Advancing a Multimethod Approach to Social Complexity*. Routledge. https://doi.org/10.1080/13645579.2025.2564177.

Powell, Copestake, & Remnant (2024). *Causal Mapping for Evaluators*. https://doi.org/10.1177/13563890231196601.

🌻 We can reason about causal maps using a logic of evidence

20 Sep 2025

> From (Powell et al. 2024)

Evaluators can break the Janus dilemma and make the best use of causal maps in evaluation by considering causal maps not primarily as models of either beliefs or facts but as repositories of causal evidence. We can use more-or-less explicit rules of deduction, not to make inferences about beliefs, nor directly about the world, but to organise evidence: to ask and answer questions such as:

• Is there any evidence that X influences Z?

• . . . directly, or indirectly?

• . . . if so, how much?

• Is there more or less evidence for any path from X to Z compared to any path from W to Z?

• How many sources mentioned a path from X to Z?

• . . . of these, how many sources were reliable?

We also argue that this is a good way of understanding what evaluators are already doing: gathering and assembling data from different sources about causal connections in order to weigh up the evidence for pathways of particular interest, like the pathways from an intervention to an outcome.

References

Powell, Copestake, & Remnant (2024). *Causal Mapping for Evaluators*. https://doi.org/10.1177/13563890231196601.

🌻 Causal maps are knowledge graphs, but with wings

26 Aug 2025

What is a Knowledge Graph? 🧠

- A knowledge graph is like a **giant mind map for a computer**. It stores information not as text in a document, but as a network of interconnected facts.
- It's built from two main things: **entities** (the "nodes," representing real-world objects, people, or concepts like "Paris" or "Photosynthesis") and **relationships** (the "edges," describing how these entities are connected, like "is the capital of" or "is a process in").
- A single fact has three parts: **(Subject) —- [Relationship] —-> (Object)**. For example: `(Marie Curie) --- [discovered] ---> (Radium)`.
- Why are knowledge graphs specially useful in the age of AI? 💡
- **They create structure from chaos.** AI can read through millions of pages of unstructured text (like news articles or scientific papers) and pull out these factual triplets. This turns a messy sea of words into an organized, queryable database of knowledge.
- **They enable smarter searching and reasoning.** Instead of just searching for keywords, you can ask complex questions that require understanding the relationships between things. For example, "Which scientists who won a Nobel Prize also discovered an element?" A computer can navigate the graph's connections to find the answer.
- **They provide essential context.** A knowledge graph helps an AI understand that "Apple" in a tech article is a company linked to "Steve Jobs," not the fruit. By looking at its connections, the AI gets the right context, which is crucial for accurate understanding and analysis.

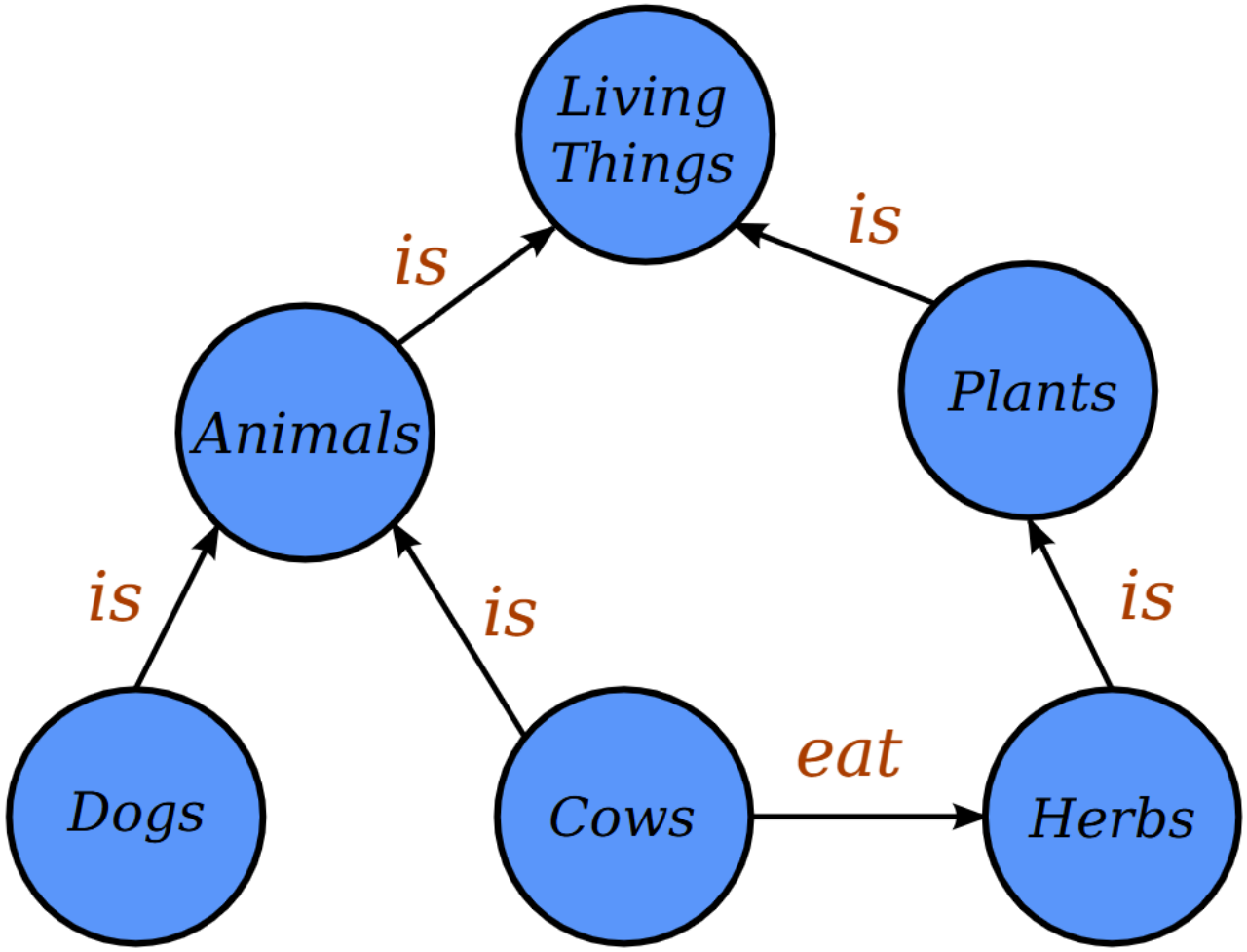


Image from Wikipedia by Jayarathina - Own work, CC BY-SA 4.0.

Why are Knowledge Graphs (KGs) so useful?

A major benefit of KGs is we can then apply network logic like transitivity rules to answer meaningful questions. For example, if the relation is "works in the same company as", then if we know

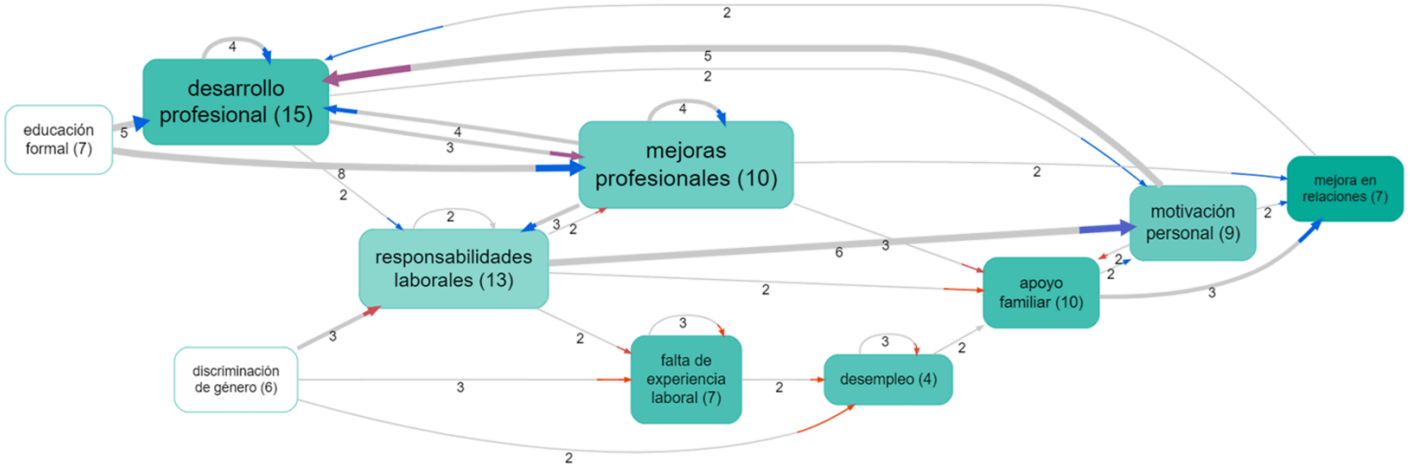
- **A is related to B**
- and
- **B is related to C,**
- then we can conclude
- **A is related to C (A is in the same company as C).**

Challenges with general-purpose knowledge graphs

The trick in **constructing** knowledge graphs is to know what relationship(s) to look for. "belongs to? " "is capital of?" "challenges/undermines?" This can be very difficult to decide. on the fly.

Using network logic to answer queries can be difficult where each different type of relationship may have its own logic. It can be very tricky (though potentially rewarding and useful) to design custom queries to answer specific questions.

Causal mapping gives knowledge graphs wings



A causal map is just a knowledge graph in which there is only one kind of relation: "causes" or "influences". This means:

- It is **much easier to scan and process text data** as we already know what we are looking for.
- We focus on primarily on **exactly the kind of information which is useful for monitoring and evaluation**: what influences what?
- We can **make use of pre-existing logic and queries to help answer common evaluation questions** almost "out of the box". Here we have a whole presentation on *Questions you can answer with causal mapping* which gives plenty of suggestions.

Can only doing causal mapping answer all the questions you might want to ask about a text? Of course not. But it can help answer a lot of the most interesting and important ones.

What about social network analysis?

Yes, social networks can also be constructed as knowledge graphs with just one (or a small number of) relationships, such as "works with".

So can we use causal mapping tools to construct general network graphs?

You might ask if the reverse is also true: can you use causal mapping software like Causal Map to also do your AI-supported knowledge graphing for you? The answer is yes!

You can code any type of link, not just causal, and you can also guide the AI to do this too.

Social Network Analysis — SNA

Quality assurance at each step of the causal coding workflow

29 May 2026

Summary: Quality assurance at each step

In the qualitative space, evaluators have many tools and approaches for reaching robust and rigorous conclusions about causal influences on an outcome of interest, perhaps as the operation of a mechanism. And evaluators are increasingly interested in causal *pathways*: multiple, multi-step, perhaps surprising paths along which influence is passed. How can we reach robust and rigorous conclusions specifically about influences *along* causal *pathways*? This briefing paper claims that *causal mapping* has a long tradition of this kind of thinking. In particular we point to some old and new features within our causal mapping app, Causal Map 4, which can help with this task.

This paper is the quality-assurance (QA) companion to the coding workflow. That paper sets out *what you do* at each step. This paper follows the same steps and asks, at each one, how to reach robust and rigorous conclusions. It is built on top of the workflow paper: where that paper describes a tool, we do not describe it again here, we discuss how to use it for QA.

Especially now that AI lets us scale a single project to tens or hundreds of thousands of causal claims, the gap between "we have many claims" and "we have warranted conclusions" is easy to underestimate. Practitioners need practical and transparent ways to get from a claim to a judgement, without overclaiming what the claims can bear.

From claims to conclusions

Causal mapping, as we practise it, is not a method of causal inference. The fact that twenty people, or twenty thousand, claim that X influences Y does not on its own warrant the conclusion that X really does influence Y. One job of causal mapping is to assemble the claims **so that an evaluator or researcher can make a judgement**. The app does not make that judgement for them. It is a preparatory step which is useful for almost any evaluation approach but especially for theory-based approaches like contribution analysis.

The longer argument for this conservative stance is in our companion paper on minimalist coding and here; see also Powell et al. (2024); Powell et al. (2023).

The Causal Map app helps at several moments in the quality assurance task.

A note on "evidence"

We have been criticised for calling the mass of causal claims "evidence": a claim is not really evidence until it has been weighed against something. But Thomas Schwandt disagreed, defining evidence as "information that has a bearing on determining the validity of a claim" — not as something which is already declared to be valid. We mostly go with Thomas. Moving from coded claims to warranted conclusions is exactly the Rubicon this paper is about. When we use "evidence" loosely, we mean only the body of claims that the evaluator can take into account, not that it has already been judged to be of any particular quality.

We have always assumed that evaluators and researchers using causal mapping and Causal Map will be doing serious quality assurance when crossing the Rubicon from claims to conclusions, but this is the first time that we have tried to address this task in more detail and point out how the Causal Map app can help with quality assurance (QA).

> Sidebar: This is separate from the way causal inference is done specifically in the Qualitative Impact Protocol (QuIP) Copestake et al. (2019) — although QuIP projects often use causal mapping, they have a more specialised and specific set of supports for causal inference.

Solving problems by breaking them down into smaller pieces

Evaluators have primarily addressed the problem of making judgements about causal influences a practical but synthetic problem of making judgements about a *contribution to an outcome*, a judgement which may in fact be about a single causal link or about a pathway or mechanism. So Outcome Harvesting for example often involves making holistic judgements about some kind of path or mechanism from intervention to outcome which is primarily presented as a single problem of "intervention influences outcome?", even though that "mechanism" may have multiple parts. (Of course, mechanisms are fractal.)

From a causal mapping perspective, it gives us a slight headache when evaluators talk about the robustness of evidence for the "causal link" or even "mechanism" from an intervention to an outcome. This holistic perspective, reducing a network of causal pathways to a single link is useful, in fact essential — it is the last of our steps, but it can gloss over a whole preceding nest of problems within the articulated causal pathways.

In this paper we break that holistic task down across the steps of the workflow.

Causal mapping provides a general, articulated framework to assemble (and then make judgements about) not only individual links (or a single bundle of links) but then about individual links combined into a pathway or network, beginning or ending with any kind of factor, not just outcomes/interventions.

This addresses the formal problem about how causal influences might or might not operate transitively down a causal pathway (if B influences C, and C influences D, does B influence D?).

But there is another formal (and practical) problem about how/if/whether our assessment of the **quality** of individual claims or bundles of claims can be assembled into an assessment of the **quality** of the evidence for a *pathway*: (if we have a validated claim that B influences C, and a validated claim that C influences D, when/how do we have a validated claim that B influence D?)

Quantitative approaches sometimes suggest that they warrant moving from data to evaluative conclusions without any "human in the loop". But at least in the qualitative world, an evaluator or evaluation team has to take responsibility for any conclusions drawn from data — especially, but not only, in the case of causal inference. All sciences help and inspire us to break problems down into smaller, reusable pieces and recombine to get the final answer. That's what this working paper is about. But however we reassemble our conclusion, we can never rely purely on the algorithm. There is a final holistic judgement to be made, even if it is just the judgement "We paid for an expensive RCT, I trust those guys, let's just publish whatever they say".

Quality assurance at each step

The steps are not really a simple sequence, and several may be revisited. Only the last is required. Most projects use several, or other overlapping approaches. Steps 1 and 2 of the workflow, planning and data gathering, carry their own quality questions; here we pick up at the codebook.

QA at Step 3: managing the codebook

See Step 3 for the mechanics.

This step may be revisited multiple times. You might start from a Zero Codebook, simply free-coding whatever you see, and you might later revise that codebook one or more times; or you might start from a more-or-less fixed codebook.

Quality Assurance in this step means asking questions like:

- Are my labels consistent? Right level of granularity?
- Whose world view do they reflect?

QA at Step 4: coding individual links

See Step 4 for the mechanics.

The most important moment for quality assurance is at the time when links are originally coded.

The two things you most want to maximise are Precision (are the links accurately coded?) and Recall, aka Coverage (did we miss any links?). The whole apparatus of coding-style choice, chunk and sample strategy, instruction-writing, and iterations in Step 4 is in the service of these two metrics. The main QA discipline here is the iterative one: test your instruction on a small, varied sample, work out specifically what is causing any problems with accuracy or coverage, tweak, and try again.

A second non-negotiable QA discipline is insisting on a verbatim quote for each and every link. Without it you are no longer showing your working, and as evaluators you cannot really justify the conclusions you draw.

QA at Step 5: checking individual links

See Step 5 for the mechanics.

In spite of all this effort, and whether you have been coding with AI or doing it yourself, there will still be some mistakes.

The first move is to tag a doubtful or surprising claim, so you can later filter such links in or out. Beyond tags, the **conviction** and **strength** columns let you record, respectively, how sure the source sounds and whether they explicitly call the influence strong or weak.

Two cautions matter for QA here. First, do not read these scales as ordinal (small/medium/large; 1/2/3). They rest on the idea that the default claim is unmarked or neutral, which is not the same as "middling". **The fact that most people do not mention the strength of a causal link when talking about it does not mean they think the links were of "medium" strength.** It just means it did not occur to them to think about or mention the strength, or that the idea of strength is not even useful or applicable in this case. Second, a conviction score is a coding of how confident the *source* sounds, not a coding of the causal strength of the link itself. Keeping that distinction clear is part of not overclaiming what the evidence says.

Source-level columns (for example distinguishing reliable from unreliable sources) feed the same QA purpose: because every link belongs to a source, source reliability becomes available for each link and can be filtered on.

Source-level checks

It can also be useful to view the links just from one source to see if they make sense and are consistent. This can involve checking the individual bundles of links between individual pairs of factors — are they consistent? See the next step.

QA at Step 6: the bundle assessment

See Step 6 for the mechanics. This step warrants its own paper; see Assessing quality or robustness of evidence for a causal link based on a bundle of coterminal causal claims for the detail.

This is the core QA move, and you should do it whether or not you use the app's formal feature for it. Look at each bundle, the claims about one link, with their context, metadata and the link-level judgements from Step 5, and weigh whether the evidence is enough to vouch for the connection. You can leave it there, having weighed the bundles by eye. Or you can record the verdict by collapsing the bundle into a single "assessed link": the underlying claims are not deleted, and a switch shows either the assessed links or the unassessed bundles, never both. Some bundles will earn no assessed link, because the evidence is too thin.

If you formalise it this way, two features keep it auditable rather than arbitrary. First, bundle-level summaries of the link-level judgements from Step 5 (for example, "in most cases conviction was neutral, with a few sources emphasising they were sure") give the human judgement a backdrop and a filter. Second, the app will not let you create assessed links, manually or with AI, until you have written your criteria into a rubric or prompt sub-panel. This is on purpose: the criteria for crossing this part of the Rubicon have to be written down. The rubric might be a five-level scale like the one Jewlya Lynn and colleagues used in their fishing industry retrospective (Lynn 2025), or just yes/no, or several dimensions like "confidence" and "degree of triangulation".

Either way, formal or by eye, the move is from a mass of raw claims to a smaller set you are willing to vouch for: a much cleaner basis for argument. A typical project might go from 1000 raw claims to 30 bundles to 25 assessed links.

Rubrics (at this step but also in steps 7, 8 and 9) can include these three criteria (Aston & Apgar (2023)):

- Plausibility (does evidence make a convincing case for the model's contribution, accounting for alternative explanations?);
- Uniqueness (does evidence point specifically to this model's practices rather than factors that would have produced similar outcomes regardless?); and
- Triangulation (are claims supported by multiple independent sources, including grantee perspectives?).

QA at Step 7: pathways and the transitivity trap

See Step 7 for the mechanics.

Even when each link, or each assessed link, is now well grounded, your work is not finished. The question of this step is how to validate claims for the *transitivity* of causation: how do we get from grounded single links to a grounded multi-step pathway?

From "A influenced B" and "B influenced C" you cannot in general conclude "A influenced C", because the contexts in which each step holds may not overlap. This is The transitivity trap, the single most important challenge for any approach that uses directed network diagrams. The canonical failure mode is: source 1 says `A -> B`, source 2 says `B -> C`, and we mistakenly conclude that anyone told a coherent story `A -> B -> C`.

Path tracing alone does not protect you from this: it shows every link on a route between your two factors, across all sources, and is easy to misread as a story someone actually told. Source Tracing is the conservative QA move. It keeps only sources that have pathways all the way from A to C, so every link on the map is part of at least one complete story told by at least one source. The app then lets you review the evidence source by source and judge whether each respondent's account is internally coherent. This is also where the **quality**-of-pathway question gets its practical answer: a pathway is only as well warranted as the within-source narratives that support it end to end.

If you have already run a bundle assessment, there is a QA trade-off in the choice to source-trace on the assessed links (clean source and citation counts, but no direct view of the quotes) or on the unassessed ones (the quotes, but a busier map). In practice you may want both, in different views.

QA at Step 8: value, relative contribution and alternative explanations

See Step 8 for the mechanics.

Judging value and relative contribution, and comparing with alternative explanations, are central (overlapping but distinct) questions in evaluation which have been really extensively covered, not least by John Mayne (2019); for that reason we won't deal with them much here, but QuIP has a lot to say about value, and see Powell (2019). From a QA point of view, the discipline is to compare the influence you care about against rival explanations on the same map, rather than examining it in isolation. See Counting and comparing influences for an approach using path/source tracing.

QA at Step 9: holistic judgement

See Step 9 for the mechanics.

Finally, you want to draw a conclusion. You have done some or all of the other steps, checked the individual causal claims, assessed the robustness of co-terminal link bundles, traced paths of influence, compared influences and alternative explanations, and finally you want to at least eyeball all the evidence again and draw a valid conclusion. But "all the evidence" might be a massive corpus. Behind a single map there are still maybe hundreds of causal claims with their associated quotes and context. Does the overall claim still make sense? Can we be sure that the links in all the pathways all belong to the same context?

The AI vignette feature can be tasked with exactly these quality questions, for example: is each link really part of a coherent, complete and consistent story from source factor (e.g. Intervention) to target factor (e.g. Outcome)? A common use is a commentary on the pathways from an intervention to a chosen outcome from the perspective of individual sources, discussing how coherent each source's story is. The AI is doing nothing more than a careful reader could do given the same inputs, and the patience to examine the quotes behind each link, so treat its draft as a starting point for your own judgement and edit it.

What the app does not do

At no point does the Causal Map app move on its own from claims to facts. Causal mapping as we see it is still, on its own, not a method of causal inference but more of a way to *identify and organise the evidence* in order for the evaluator or researcher to make causal inferences, especially when assisting established methods like Contribution Analysis or QuIP. Still, in the past we have perhaps not done enough to say how exactly to do this or to make it easier to do. This paper hopes to redress that.

The warranting is always the evaluator's. We provide structures (tags, columns, the assessed-link switch, source tracing, vignettes) that make warranting easier, more transparent, and more auditable. We do not provide an engine that turns "twenty people said so" into "therefore it is so".

The opposite design, in which an algorithm rules on causal truth from coded text, would either smuggle in strong assumptions about variables and functional forms (which we argue against in Our approach is minimalist — we do not code the strength of a link and at length in our minimalist coding paper) or conflate evidence volume with effect size, which Causal Map has always been at pains to avoid. As we put it elsewhere, "a coded link is first and foremost 'there is evidence that a source claims X influenced Y', not a system model with weights or effect sizes" (Powell et al. 2024).

How this relates to other strength-of-evidence methods

The bundle and pathway judgements above are our practical contribution to a question the wider causal pathways field takes seriously: how do you assess the strength of evidence behind a causal claim (Apgar & Aston 2025)? Some methods build the test in. Process tracing weighs each link with hoop and smoking-gun tests (Befani & Stedman-Bryce 2017; Collier 2011); contribution analysis builds and tests a contribution story (Mayne 2012). Where a method has no such test built in, a written rubric does the same job, agreed in advance of the evidence, as in the CLARISSA programme's quality-of-evidence rubrics (Apgar 2024) and Jewlya Lynn's seafood retrospective (Lynn 2025). The rubric in Step 6 is exactly this device. None of these removes the final evaluative judgement; they make it transparent and auditable.

None of this is causal inference

None of this is causal inference in a statistical sense. It is a disciplined way to assemble evidence, weigh it transparently, and reach conclusions that you can defend.

This all works, we use it every day in our consultancy work at Causal Map Ltd., but it is still also evolving every day, so if you are interested in going on this journey with us, do get in touch.

Footnote: The same QA problematic and logic applies even when the links are not strictly causal: in social network analysis or other map-based work, you may still want to go from a mass of raw claims to a smaller set of checked or verified links, even though the links are about relationships rather than causation. Causal Map can do this too, and the mechanics described in the workflow paper work in the same way, though our main focus here is specifically on *causal* links.

Related

- A workflow for causal coding with and without AI: the workflow this paper assures
- Assessing quality or robustness of evidence for a causal link based on a bundle of coterminal causal claims: detail on the bundle assessment step
- Minimalist coding for causal mapping: the coding stance
- Our approach clearly distinguishes evidence from facts and does not automatically warrant causal inferences
- The transitivity trap
- The most controversial feature of causal maps is transitivity
- Just add rigour Three do's and don'ts

References

Apgar (2024). *A PARTICIPATORY APPROACH TO EXPLORING CAUSAL PATHWAYS Experience from the CLARISSA Programme July 2024*. https://www.causalpathways.org/_files/ugd/5a867c_7c84e6119d1245059c17fd4cb65d3422.pdf?utm_campaign=e55496ed-f18a-4cd0-a785-8da9f92ca069&utm_source=so&utm_medium=mail&cid=3922b538-ab12-430e-b31a-927bc02905ab.

Apgar, & Aston (2025). *How Do We Define and Support Quality and Rigor in Causal Pathways Evaluation?*.

Aston, & Apgar (2023). *Quality of Evidence Rubrics for Single Cases*.

Befani, & Stedman-Bryce (2017). *Process Tracing and Bayesian Updating for Impact Evaluation*. http://dx.doi.org/10.1177/1356389016654584.

Collier (2011). *Understanding Process Tracing*. https://doi.org/10.1017/S1049096511001429.

Copestake, Morsink, & Remnant (2019). *Attributing Development Impact: The Qualitative Impact Protocol Case Book*. March 21, Online.

Lynn (2025). *HU Seafood Retrospective*. https://www.policysolve.com/resources/retrospective.

Mayne (2012). *Making Causal Claims*.

Mayne (2019). *Assessing the Relative Importance of Causal Factors*.

Powell (2019). *Theories of Change: Making Value Explicit*.

Powell, Larquemin, Copestake, Remnant, & Avard (2023). *Does Our Theory Match Your Theory? Theories of Change and Causal Maps in Ghana*. In *Strategic Thinking, Design and the Theory of Change. A Framework for Designing Impactful and Transformational Social Interventions*.

Powell, Copestake, & Remnant (2024). *Causal Mapping for Evaluators*. https://doi.org/10.1177/13563890231196601.

🌻 Assessing quality or robustness of evidence for a causal link based on a bundle of coterminal causal claims

11 Dec 2025

This post gives more details on one of the key moments for Quality Assurance in causal mapping, the bundle assessment, which is Step 6 of the coding workflow and is discussed in the companion quality assurance paper.

Causal mapping is a way of analysing qualitative data, what people say in interviews, focus groups, reports or any written source, when you want to understand what they think causes what. This post is about a problem that surfaces once you have one of these maps in front of you. Different sources, or different parts of a single source, often make similar causal claims from the same X to the same Y, sometimes reinforcing each other, sometimes pulling in opposite directions. We call that group a *bundle* of links. How confident should we be, across the bundle as a whole, that X really does influence Y? Up to now we have largely left that judgement to the analyst's eye. Here we sketch a more systematic way to record it.

I was inspired by a recent talk by Jewlya Lynn about a causal mapping evaluation she and her team conducted Lynn (2025), in particular how they made evaluative judgments about the overall strength or robustness of evidence for the claim that one thing influenced another. We have been working on a similar idea and Jewlya's excellent report has encouraged us to move it forward.

So what is the problem exactly?

After your initial causal mapping, you will usually end up with multiple causal claims for a given single link or path from X to Y.

The Causal Map app partly grew out of our causal mapping work with QuIP evaluations. QuIP has its own way of dealing with quality and robustness of evidence. And it works mostly with relatively homogenous data sets (similar interviews with sets of similar respondents) so "number of links" can be a ballpark proxy for "strength of connection". But when working with heterogeneous sources of evidence, this does not work. In the past we have said that it is up to the analyst to look at the claims and make their own assessment about the strength of a claim that one thing influenced another, perhaps via multiple steps. But this is quite a big ask for the analyst to look at all the information from all the causal claims each time.

Just a reminder about our terminology around "links:"

Relevant page: Bundle of Links — definition ▸

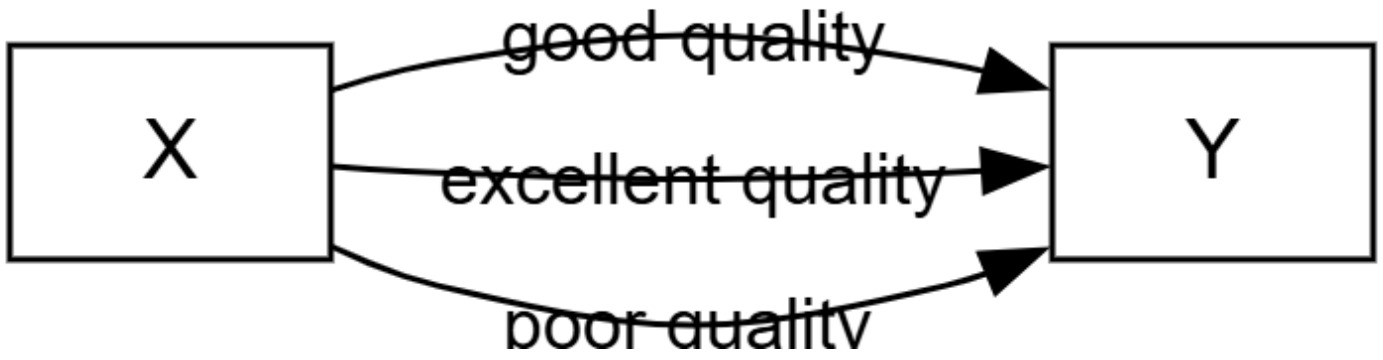
Normally, not one of these possibly hundreds or thousands of causal claims, grouped into many bundles, is incontrovertible. Sometimes we call each of these claims "evidence" but only in a weak sense of "something we could take into consideration when weighing up the validity of the claim that X causally influenced Y". Usually these links — singly or as part of bundles — have not yet passed any test at this stage or been compared to any standard.

So how can we submit our links and bundles of links to this kind of assessment?

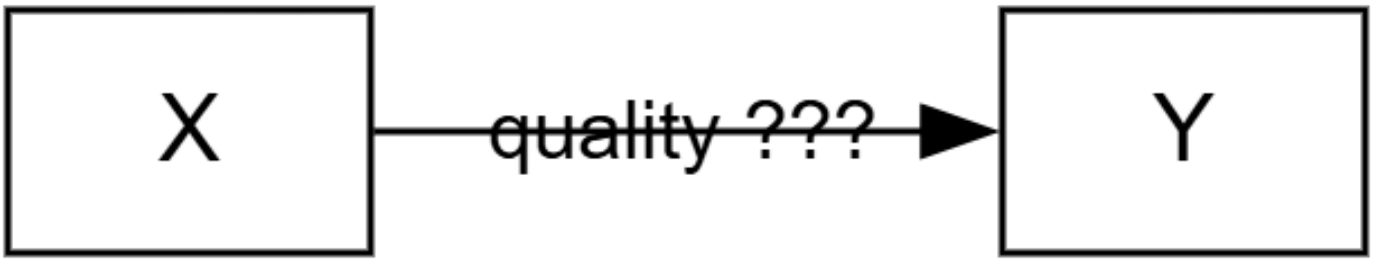
What we can already do using existing Causal Map functionality is filter individual links. For example, we can (during or after initial coding) just add a tag `doubtful` for doubtful claims and exclude these from most or all visualisations using a "include tags" or "exclude tags" filter. But these judgements are based only on individual links. In actual evaluation and research practice there is a strong case for making more global assessments about the quality or robustness of the complete set of links within the bundle. So we add a new layer to the workflow and transform a map consisting of very many individual links within bundles into a map consisting just of a single links representing each bundle, each of which carries a global assessment of the quality or the business of the evidence within the original, unassessed, bundle.

This new layer can also take advantage of bundle-level summaries of judgements made at the level of individual links. When you look at a bundle of claims for X influences Y, the Causal Map app now summarises the distribution in a sub-panel of the Assessment panel: for example, reporting that in most cases conviction was neutral, with a few sources emphasising they were sure. This is helpful both as a backdrop for human judgement and as a filter (for example, exclude links where the source said they were uncertain). See Coding with and using link metadata for the mechanics.

The workflow goes from something like this (3 unassessed links):



... to something like this (1 assessed link) ....



So all four links are present in the database, but we always show either the assessed links or the unassessed links.

In more detail:

1. Analyst finishes coding the map (whether human coding, AI coding or some combination).
2. Analyst fixes on a set of bundles to take seriously: those that survive any filters, perhaps after zooming to a higher level of the coding hierarchy or restricting to particular sources or subgroups. There might be five, fifty, or a hundred such bundles. This is the data that the rest of the analysis rests on.
3. Analyst considers each bundle (including bundles that might contain only one link) and judges the quality or strength of the evidence, by hand or with AI assistance plus review.
4. Analyst collapses the bundle into a single new "assessed link" carrying one or more quality assessments. The assessed link is a new type of object in the links database. By default it inherits the citation count and source count of the underlying bundle, and can carry additional scores from custom columns. In the simplest case the assessment could be:
    1. Robust yes/no
    2. or: Robustness 1-5
5. Some bundles will not yield an assessed link at all, because the evidence is too thin. You can either skip the bundle (creating no assessed link) or create one with a custom column "Passed?" set to "Fail".
6. For each bundle, then, there is a set of (usually multiple) unassessed links and at most one assessed link. From now on you will normally view either the assessed links or (less often) the unassessed links.
7. Analyst uses the existing links table, pivot tables and/or map formatting to display these assessed links. Most obviously, they can show overall maps with only the quality-assessed links, formatted according to quality.

It's saying this:

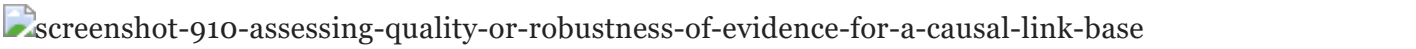
> I the analyst have looked at this chunk of quotes and contexts for this one bundle and I vouch for the judgement that it's enough to say yes there's something real going on here.

You can page through the bundles by hand, or you can let the AI do a first pass against a rubric you supply, and then review its work. The app will not let you create assessed links, manually or with AI, until you have written your criteria into a rubric or prompt sub-panel. This is on purpose.

Unassessed view: Links table. Many links in each bundle.

*Bookmark #1485 2026-04-29 07:41_*

Assessed view: Links table. One link for each unassessed bundle. Only 6 links in total.



*Bookmark #1484 2026-04-29 07:38_*

The result is a parallel map. The unassessed claims remain in the database, but a switch in the app lets you view only the assessed links (or only the unassessed). A typical project might go from 1000 raw claims to 500 filtered claims in 30 bundles to 25 assessed links. You can use the newish "Map Custom Columns" filter to apply custom formatting to your final maps by source count, citation count, or any custom score (degree of triangulation, for example). The simpler, assessed map gives you a cleaner basis for argument than the raw claims.

The big question is of course, what criteria should we use to make these robustness or quality assessments. The answer has to be based on use: what are we actually trying to do here? We might for example want to focus on the quantity or quality or robustness of the evidence taken as a whole.

In Lynn (2025), here is the matrix which Jewlya Lynn and colleagues used.

Lynn (2025)

My feeling is that, like Jewlya, we probably want to collapse all this information to just a single dimension, perhaps 1-5. Or you might want to keep multiple dimensions, for example "confidence" and "degree of triangulation". The decision is yours.

It is also possible to summarise this information not in numerical or binary form at all but simply as text judgement like "Seems solid but not really sure there is enough evidence specifically that this works for children too". This means writing into one or more text "memo" columns for the assessed links.

Noting absences

Normally you'd add one assessed link for each bundle of Unassessed links, and in the UI you have a switch or filter to show either the one or the other, which then determines what you see in the outputs (maps and tables).

But there is nothing to stop you putting in an Assessed link for a bundle for which there are no claims at all. That is something which has been hard in Causal Map up to now.

Something else to think about

We add these "assessed" links on a per-bundle basis. But the bundles might not contain the original cause and effect labels because we can add them also for filtered labels, e.g. after zooming or after applying soft recoding to the original labels. That means you might have a project containing a set of assessed links which are only indirectly based on the original codings.

References

Lynn (2025). *HU Seafood Retrospective*. https://www.policysolve.com/resources/retrospective.

🌻 The product of (causal) qualitative coding can be a model you can query

The result of qualitative coding of texts: is it a model?

Every type of Qualitative Data Analysis (QDA) is a bit different....

1. The most important result of coding might be a **deeper understanding of the text**. Something the researchers have in their heads. Perhaps something shared within a group. They can answer new questions about the data and write summaries of it from a new angle. In a sense you could say they have more or less formally developed a **model** of the data or even a theory about it, or, in the sense of Grounded Theory, a theory which goes beyond that data but is partly inspired by it. If you yourself read some of these different outputs and engage with them, maybe you yourself can start to build such a model in your head.

2. Apart from the intangible products in researchers' heads there are also **tangible products such as research reports**. Different schools of QDA give different importance to tangible as opposed to intangible outputs.

3. Thirdly, there may be **tabular outputs** such as tables of coding frequencies, cross-tabulations, etc. The definitive set of tables can be queried to answer additional questions like "how many women who mentioned working from home also see television as a an outdated medium". These kinds of tabular outputs can be quite useful to answer different questions.

So when you do an evaluation, what's the product? What do you get?

Obviously, you can think of the evaluation report, which might answer predetermined questions, but it may also include material that goes beyond the specific questions we were tasked with answering—for example, to address unanticipated issues or simply to describe or contextualize. Another important output is relational: hopefully, people have come together in a way that helps to expand learning and perhaps develop projects or relationships.

But today I want to talk about something different.

A statistical analysis not only answers questions but gives you the whole model — can a qual analysis do that too?

When you do quantitative research, you might have specific research questions, but often one of the major outputs is a statistical model of the phenomenon. (There might be an effort to go beyond the data

and hope that the model generalizes more widely, but that's not my focus here.) In the simplest case, the model might represent a suspected causal relationship—say, between the amount of screen time in the evening and difficulty falling asleep.

At the very least, the model allows us to look at a case in the dataset and say: on this day, this person looked at a screen for three hours and rated, let's say, a difficulty of four out of five falling asleep on some self-rating scale. Because we have the model, **we can explain that**: yes, this is quite a high level of difficulty, and it's explained at least partly by a high level of screen time, at least in this individual case. The model might also enable us to **make predictions**, like: people who, at least in this context, spend more than three hours on screens in the evening are, on average, going to experience a higher level of difficulty falling asleep.

A more sophisticated model will probably capture more variables, and many models—like directed acyclic graphs—link up these kinds of connections into a causal network, so you can explain or even predict how tweaking one variable will affect another variable downstream of it. Of course, there are other kinds of statistical models apart from causal models, but if you're reading this, you love causal models, don't you?

There's a relatively small but extremely well-funded section of evaluation activity based around this kind of statistical causal model, with randomized control trials (RCTs) as one facet. Ideally, an RCT is tasked not primarily with producing a model, but with answering a specific question, like: which is the better of these two interventions? Or: does this treatment work better than a placebo? But these are calculations conducted on the underlying model, which, from the point of view of workflow, is the major output of the work.

To generalise or not to generalise

There are two ways you might want to use that kind of model.

- One is to answer further questions about the same dataset—for example, to ask whether a particular subset (say, people over 70) differ in how screen time influences difficulty falling asleep, compared to other subgroups.
- If it's a sophisticated model, it might allow us to *generalize beyond* this specific context, perhaps by including more general variables like attention style or eye movement speed, which might help explain and predict behavior in other contexts.

Most quantitative researchers make a big deal about generalizing the model beyond the specific use case or context. In fact, the whole point of the study is normally to do that, and there's a whole armoury of tools, concepts, and arguments about how and under what conditions you can generalize a model to other people, other years, or even other countries with different kinds of screens, etc.

Can qualitative research do that?

The majority of evaluations aren't like that, although they might include a specific quantitative question somewhere in their terms of reference. In most cases, we think of the research output as a report in which the original (possibly modified) list of questions is answered, with additional narrative to summarize and link these sections.

Now, going beyond strictly quantitative paradigms, some evaluation projects will also include what we might call **qualitative modelling**. For example, if we're using QCA (Qualitative Comparative Analysis), apart from answering specific questions, we've likely also produced QCA-style tables, which could help us answer other questions beyond those we were actually tasked with. You might see those tables as annexes to the report. The same goes for causal loop diagrams and other techniques, which are essentially quantitative models but with a more restricted set of numbers. For example, in causal loop diagrams, we might model a variable like inflation with a number from -1 through 0 to +1, and do the same for variables like unemployment or military threats, building models of the relationships between these things using simplified numbers.

What I want to argue here is:

> any halfway decent evaluation, which at least implicitly gathers qualitative information about how things within the evaluation influence one another, can be considered as constructing a qualitative causal model. This is irrespective of the specific methods used—even if it doesn't include something explicitly called causal pathways analysis or causal mapping.

Qualitative models as products: theory, model, or something else?

In quantitative research, the idea of a "model" as a product is well established: you build a model, and then you can query it to answer new questions—even ones you didn't anticipate at the start. But what about qualitative research? Can the result of a qualitative analysis be a model in this sense, rather than just a set of answers to specific questions or a summary?

Of course (some) qualitative research produces models. Just don't call them that.

Some qualitative researchers do indeed conceptualize their results as models. For example, grounded theory often produces a theoretical model that explains the underlying processes or relationships within the data. These models can be revisited and "queried" to generate new insights beyond the initial research questions.

However, many qualitative researchers are more comfortable with the term "theory" rather than "model." "Theory" aligns more closely with the interpretive and conceptual nature of qualitative work, emphasizing explanation and understanding rather than prediction or parameterization. Still, the distinction is often

more about language than substance: both models and theories can serve as frameworks for making sense of data and for generating new questions.

How are qualitative models used?

In qualitative research, especially in fields like grounded theory and narrative inquiry, the focus is less on "prediction" and more on generating understanding or insight. Researchers talk about "theorizing" from the data—developing concepts and frameworks that explain the phenomena under study. Once a theory or model is developed, it can be revisited, interrogated, and applied to new data or different contexts. This iterative process allows for continual refinement and deeper insight.

Importantly, a qualitative model or theory can also be used to answer new questions about the same dataset. For example, after developing a grounded theory, researchers (or others) can return to the theory and use it as a lens to interpret further cases, refine concepts, or generate new insights—without having to re-examine all the original data. This practice is sometimes referred to as "secondary analysis" or "theoretical application," where the theory or model functions as a standalone analytical tool.

In causal mapping, for instance, the model might consist of a network of causal links derived from qualitative data. Even if there are no quantitative parameters on the links, the model can still be queried: "Is there evidence for a causal pathway from A to B?" or "What are the main factors influencing outcome X?" This allows the model to be used flexibly, supporting both anticipated and unanticipated lines of inquiry.

Why does this matter?

Thinking of qualitative research outputs as models (or theories) that can be queried and reused has several advantages:

- **Transparency:** It makes explicit the structure of the findings and how they relate to the data.
- **Reusability:** Others can use the model/theory to answer new questions or apply it in new contexts.
- **Iterative learning:** The model can be refined and expanded as new data or perspectives emerge.
- **Bridging paradigms:** It helps bridge the gap between qualitative and quantitative traditions, showing that both can produce structured, interrogable outputs.

In summary, while qualitative researchers may prefer the language of "theory" over "model," the idea is the same: a well-constructed qualitative analysis can produce a framework that is more than just a set of answers—it is a model of the phenomenon, one that can be queried, shared, and built upon.

Related

- chapter intro

The transitivity trap

22 Aug 2025

> From (Powell et al. 2024)

Granularity, generalisability and chunking are coding problems for causal mapping too

Transitivity is perhaps the single most important challenge for causal mapping. Consider the following example. If source P [pig farmer] states 'I received cash grant compensation for pig diseases [G], so I had more cash [C]', and source W [wheat farmer] states 'I had more cash [C], so I bought more seeds [S]', can we then deduce that pig diseases lead to more cash which leads to more seed (G à C àS), and therefore G à S (there is evidence for an indirect effect of G on S, i.e. that cash grants for pig diseases lead to people buying more seeds)?

The answer is of course that we cannot because the first part only makes sense for pig farmers, and the second part only makes sense for wheat farmers. In general, from G à C (in context P) and C à S (in context W), we can only conclude that G à S in the intersection of the contexts P and W. Correctly making inferences about indirect effects is the key benefit but also the key challenge for any approach which uses causal diagrams or maps, including quantitative approaches (Bollen 1987).

> For want of a nail the shoe was lost,
>
> For want of a shoe the horse was lost,
>
> For want of a horse the rider was lost,
>
> For want of a rider the battle war lost,
>
> For want of a battle the kingdom was lost,
>
> And all for the want of horseshoe nail.

(Thanks to Gary Goertz for remembering this one!)

Frog thinks: eating salad leads to health (less scurvy), and health (general fitness) leads to better sprinting ability, therefore if I eat this yummy lettuce – AARGH!

One of the key features of causal maps is that you can draw inferences, make deductions, from them. One of the most exciting is to be able to trace causal influences down a chain of causal links. BUT, when you are drawing conclusions from causal maps, beware of the transitivity trap:

from

> B → C

and

> C → E

we can only conclude

> B → E in the intersection of the contexts of 1 and 2

… and in general with any causal mapping, you'll never be sure that these two contexts do intersect. You actually have to look at each chain and think about it, and hope you've been told all the relevant facts.

For example:

If

> Source P [pig farmer]: I received cash grant compensation for pig diseases (G), so I had more cash (C)

and

> Source W [wheat farmer]: I had more cash (C), so I bought more seeds (S)

can we deduce

> G → C → H

and therefore

> G → S

(cash grants for pig diseases lead to people buying more seeds)?

No, we can't, because the first part only makes sense for pig farmers and the second part only makes sense for wheat farmers.

There are thousands of different kinds of transitivity trap. It isn't just a problem across subgroups of people. It can apply for example in different time frames.

If

> Child does well in year 13 (A) → Child has improved academic self-image (C)

and

> Child has improved academic self-image (C) → Child does better in year 9 (D)

can we deduce

> A → C → D

and therefore

> A → D

(child doing well in year 13 leads to child doing well in year 9)?

Of course not - even though these claims might be true of the same child. The problem arises as soon as we generalise one causal factor to apply to different contexts. We have to do this, to make useful knowledge. But there are always pitfalls too.

Not just a problem for causal mapping

This is also true, isn't it, of any synthetic research / literature review?

And in statistics, knowing the effects from B → C and C → E means you can calculate the indirect effect of B on E but not the direct effect. You have to have additional data just for that. This is one source of various so-called paradoxes in statistics.

Can we mitigate the trap with careful elicitation protocols?

Sometimes, we might know that all the information in one particular chain came from the same source, and all this information was explicitly given as a series of explanations of the factor which was initially in focus. But even here, we have to be careful. We might have to ask again, having reached the end of the chain, "did B really influence C which influenced D which influenced E? Was this all part of the same mechanism?" Are we sure we know exactly what we mean by this, and are we sure that our respondents do too?

In any case, part of the point of causal mapping is the synthetic surprises which we can discover by piecing together fragments of causal information which were *not* necessarily provided in this way. This is the situation every evaluator is in when piecing together information from, say, experts for Phase 1 and experts for Phase 2. We just always have to be aware of the transitivity trap.

Transitivity trap, or identity trap?

We can talk about the *identity trap* as more fundamental than the transitivity trap. It comes down to saying, how can you be sure that the way in which this factor is exemplified in one particular context is the same as the way that this similar seeming factor is exemplified in a different context: whether to use “the same” factor to code two different things.

References

Bollen (1987). *Total, Direct, and Indirect Effects in Structural Equation Models*. JSTOR.

Powell, Copestake, & Remnant (2024). *Causal Mapping for Evaluators*. https://doi.org/10.1177/13563890231196601.