The Clean Paper

WHAT THE PAPER SAYS · WHAT IT DOESN'T · WHY IT MATTERS

Một bằng chứng mật mã có thể che bí mật bằng cách khiến việc chứng minh rằng bộ mô phỏng không tồn tại trở nên khó

Bằng chứng zero-knowledge cho phép ai đó chứng minh một mệnh đề mà không tiết lộ witness. Lý thuyết cổ điển nói rằng, theo nghĩa đầy đủ, bạn không thể có đồng thời một thông điệp duy nhất, không thiết lập đáng tin cậy và tính đúng đắn hoàn hảo. Bài báo của Rahul Ilango không làm kết quả bất khả thi đó biến mất. Nó đổi mục tiêu: thay vì đòi hỏi một bộ mô phỏng thực sự tồn tại, nó yêu cầu rằng không hệ chứng minh được chọn nào có thể chứng minh hiệu quả rằng bộ mô phỏng không tồn tại. Dưới các giả định lớn của độ phức tạp chứng minh và mật mã học, điều đó đủ để khôi phục từng hệ quả có thể kiểm nghiệm bằng trò chơi của zero-knowledge. Mục tiêu không phải một primitive mật mã có thể cắm vào Internet để dùng ngay, mà là một cách dựa trên lý thuyết chứng minh để biến tính không chứng minh được trong toán học thành một lớp che phủ mật mã.

Written by Cinzia Vaglio · Cross-checked by Vera Rubin · Edited by Michele Renda · Figures by Nova Vela · AI-assisted, human-reviewed

Paper: *Gödel in Cryptography: Effectively Zero-Knowledge Proofs for NP with No Interaction, No Setup, and Perfect Soundness*, Rahul Ilango, FOCS 2025 / IACR ePrint 2025/1296 · DOI: 10.1109/FOCS63196.2025.00058

Mẹo ở đây không phải là chứng minh bí mật đã được che giấu

Hãy bắt đầu với phiên bản đơn giản nhất của zero-knowledge.

Alice muốn thuyết phục Bob rằng một câu đố Sudoku có lời giải. Nếu cô gửi luôn lời giải, Bob sẽ tin, nhưng câu đố cũng mất vui. Điều Alice muốn kỳ lạ hơn: một bằng chứng rằng lời giải tồn tại, nhưng không tiết lộ lời giải.

Đó là lời hứa của một **bằng chứng zero-knowledge**. Người chứng minh (*prover*, Alice) thuyết phục người xác minh (*verifier*, Bob) rằng một mệnh đề là đúng, đồng thời không tiết lộ gì ngoài chính sự thật của mệnh đề đó.

Vấn đề là lời hứa này phải trả giá. Một bằng chứng toán học thông thường có hai đặc điểm rất dễ chịu. Nó là **một thông điệp duy nhất**: bạn viết xuống, đưa cho người khác rồi rời đi. Và nó có **tính đúng đắn hoàn hảo** (*perfect soundness*): một mệnh đề sai hoàn toàn không có bằng chứng hợp lệ. Các kết quả bất khả thi cổ điển nói rằng zero-knowledge phải từ bỏ cả hai đặc điểm này — và không chỉ khi đòi hỏi chúng cùng lúc; từng đặc điểm riêng lẻ cũng đã bị loại trừ.

Thứ nhất, một bằng chứng zero-knowledge cần có đối thoại. Nếu Alice chỉ gửi một thông điệp, không có thiết lập đáng tin cậy được chuẩn bị từ trước, bảo đảm zero-knowledge sẽ sụp đổ — bất kể bạn sẵn sàng hy sinh bao nhiêu tính đúng đắn để đổi lấy nó.

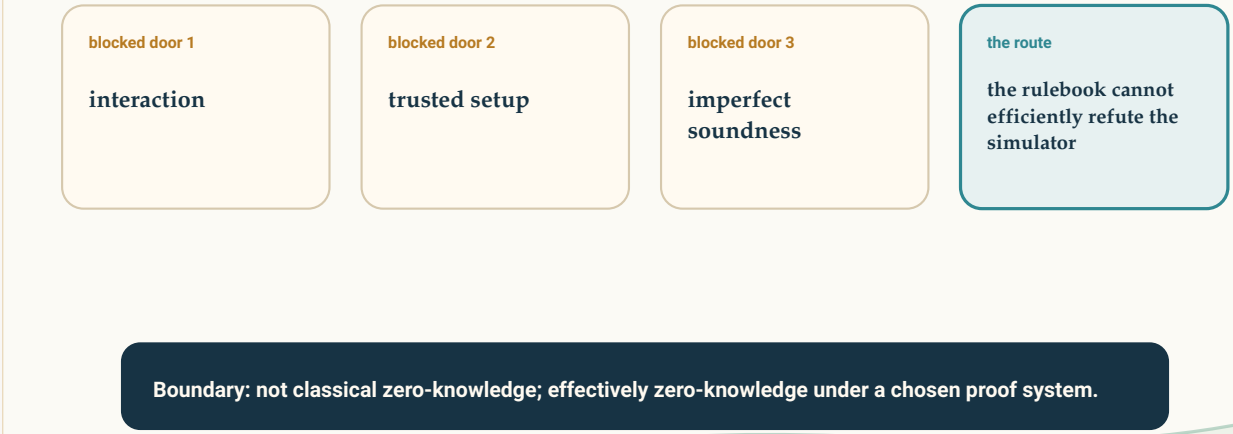
Thứ hai, một bằng chứng zero-knowledge cần chấp nhận một xác suất sai rất nhỏ. Đòi hỏi tính đúng đắn hoàn hảo hóa ra cũng âm thầm phá hủy tính tương tác: nếu người xác minh không bao giờ có thể bị đánh lừa, dù các lựa chọn ngẫu nhiên của họ là gì, thì họ cũng có thể cố định các lựa chọn đó ngay từ đầu — và khi người xác minh trở nên dự đoán được, Alice có thể trả lời mọi thứ trong một thông điệp duy nhất, chính là trường hợp đã thất bại ở trên.

Bài báo của Rahul Ilango nói về một cách đi vòng qua bức tường kép đó. Không phải bằng cách giả vờ bức tường không tồn tại, cũng không phải bằng cách tạo ra zero-knowledge cổ điển trong một bối cảnh vốn bất khả thi. Nước đi tinh tế hơn: làm yếu ý nghĩa của “không tiết lộ gì”, nhưng làm yếu theo cách vẫn bảo toàn các thuộc tính an ninh mà nhà mật mã học thực sự có thể kiểm tra.

Kết quả được gọi là **zero-knowledge hiệu dụng** (*effectively zero-knowledge*).

A proof without leaking the witness

The paper keeps the ordinary-proof shape by weakening what privacy must prove.

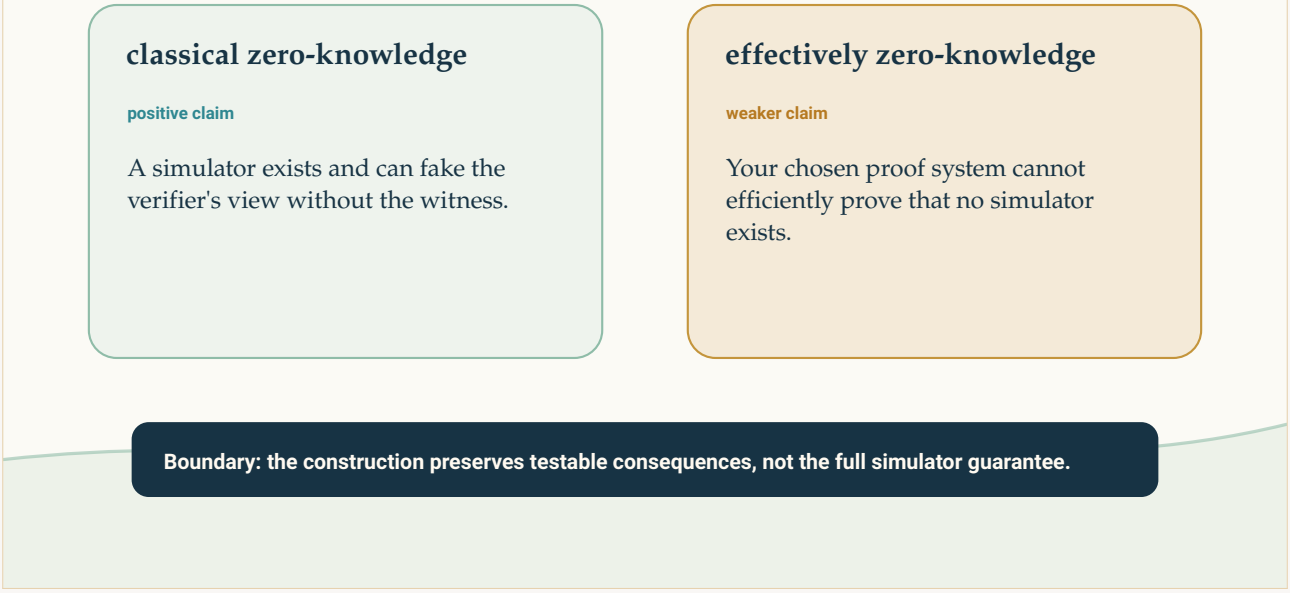


Zero-knowledge bị chặn ở ba cánh cửa — tương tác, thiết lập đáng tin cậy và tính đúng đắn không hoàn hảo. Cấu trúc của Ilango đi qua một lối khác: bộ quy tắc không thể bác bỏ bộ mô phỏng một cách hiệu quả.

Original diagram — The Clean Paper CC BY 4.0

Classical privacy vs effective privacy

The relaxation is the point: the paper changes what the privacy claim has to establish.



Zero-knowledge cổ điển hỏi liệu một bộ mô phỏng có thực sự tồn tại hay không; “zero-knowledge hiệu dụng” chỉ hỏi liệu bộ quy tắc bạn chọn có thể chứng minh hiệu quả rằng nó không thể tồn tại hay không. Chính câu hỏi

yếu hơn này cho phép cấu trúc giữ được một thông điệp, không thiết lập và tính đúng đắn hoàn hảo.

Phép thử cũ: một bộ mô phỏng tồn tại

Cách cổ điển để hình thức hóa zero-knowledge dùng một trợ thủ hư cấu gọi là **bộ mô phỏng** (*simulator*).

Ý tưởng là thế này: hãy tưởng tượng Jane, người **không** biết bí mật của Alice. Nếu Jane có thể hoàn toàn tự mình tạo ra những bằng chứng trông giống hệt những gì Bob sẽ nhận được từ Alice, thì bằng chứng của Alice đã không dạy cho Bob điều gì mới. Jane vốn đã có thể giả lập trải nghiệm đó mà không cần bí mật của Alice.

Vì vậy, zero-knowledge cổ điển đòi hỏi một bộ mô phỏng thực sự. Phải tồn tại một thuật toán hiệu quả có thể tạo ra các bằng chứng giả trông thật mà không biết bí mật — trong thuật ngữ chuyên môn là *witness*, ở đây có thể hiểu là **nhân chứng**; với Sudoku, witness đơn giản là lưới đã giải xong.

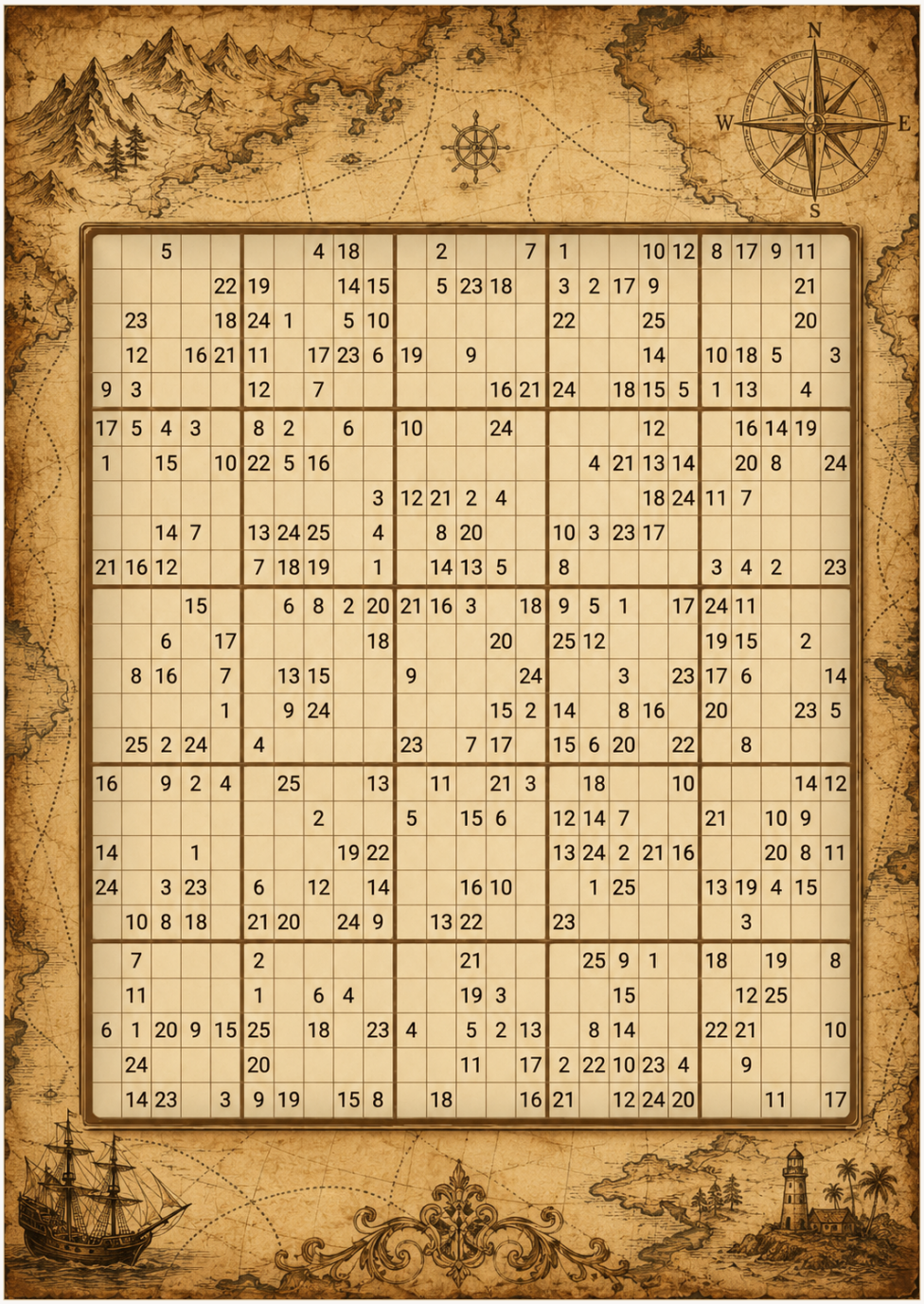
Định nghĩa đó rất mạnh, nhưng cũng chính là nơi các kết quả bất khả thi cũ phát huy tác dụng. Trực giác là thế này. Một bằng chứng thật sự không tương tác chỉ là một chuỗi ký tự. Khi Bob đã có chuỗi đó, anh ta có thể đưa nó cho người khác: anh ta đã có khả năng chứng minh mệnh đề cho người thứ ba, và điều đó nghe đã giống nhiều hơn “không học được gì”. Các định lý cổ điển biến trực giác ấy thành những kết quả bất khả thi nêu trên.

> **Ba thuộc tính mà bài báo nhất quyết giữ lại**
>
> Tiêu đề bài báo nêu ba ràng buộc:
> **Không tương tác:** Alice gửi một chuỗi bằng chứng duy nhất. Không có giao thức hỏi đáp qua lại.
> **Không thiết lập:** Alice và Bob không dựa vào một chuỗi tham chiếu chung đáng tin cậy hay nguồn ngẫu nhiên công khai nào được chuẩn bị từ trước. Nhiều hệ thống được gọi là “zero-knowledge không tương tác” vẫn dựa vào một bước thiết lập; bài báo này thực sự có nghĩa là không thiết lập.
> **Tính đúng đắn hoàn hảo:** một mệnh đề sai không có bằng chứng hợp lệ. Không phải “gần như không bao giờ được chấp nhận”; mà là không tồn tại bằng chứng hợp lệ.
> Ba thuộc tính này chính là những gì một bằng chứng toán học viết trên giấy thông thường có — và, như đã giải thích ở trên, zero-knowledge cổ điển không thể giữ chúng.

Một phiên bản MegaSudoku để cảm nhận sự khác biệt

Đây là một cách cố ý đơn giản hóa để cảm nhận sự khác biệt.

Đừng dùng Sudoku 9×9 thông thường cho phần nghiêm túc của phép so sánh. Nó quá nhỏ và quá hữu hạn: máy tính có thể đơn giản giải nó, hoặc chứng minh rằng nó không có lời giải. Thay vào đó, hãy tưởng tượng một họ câu đố **MegaSudoku(n)**. Mở rộng quy tắc quen thuộc: chọn kích thước khối *n*, đặt *N = n^2*, rồi tạo một lưới $N \times N$ chia thành các khối $n \times n$, với *N* ký hiệu. Sudoku thông thường chỉ là trường hợp bé xíu *n = 3*, *N = 9*: lưới 9×9, các khối 3×3 và chín ký hiệu. Câu chuyện về độ phức tạp của bằng chứng chỉ thực sự bắt đầu khi *n* được phép tăng và khi lưới có thể mang thêm các gadget khiến nó hành xử như một công thức SAT khoác áo Sudoku. Một công thức SAT đơn giản là một danh sách các ràng buộc có/không: liệu ta có thể gán giá trị đúng/sai cho các biến sao cho mọi ràng buộc đều được thỏa mãn hay không?



Một Sudoku 25×25: các quy tắc của nó có thể được kiểm tra mà không cần tiết lộ lưới hoàn chỉnh — một hình ảnh thay thế cho bằng chứng xác minh một lời giải ẩn, tức witness.

Sudoku và SAT: cùng một câu đố trong hai bộ trang phục

Khẳng định rằng Sudoku có thể "hành xử như một công thức SAT" không chỉ là ẩn dụ. Ta có thể dịch theo cả hai chiều, và chiều dễ hơn có thể viết ra đầy đủ.

Từ Sudoku sang SAT. SAT chỉ nói bằng đúng/sai, vì vậy hãy tạo một biến Boolean cho mỗi bộ ba (hàng, cột, giá trị): *x(r,c,v)* có nghĩa là "ô ở hàng *r*, cột *c* chứa giá trị *v*". Một

Sudoku 4×4 (khối 2×2, các giá trị 1–4) cần 4·4·4 = 64 biến; Sudoku 9×9 cổ điển cần 729. Sau đó, mỗi quy tắc Sudoku trở thành một nhóm mệnh đề. (Một mệnh đề là phép OR của các biến hoặc phủ định của chúng; toàn bộ công thức là phép AND của tất cả các mệnh đề.)
Mỗi ô chứa ít nhất một giá trị — một mệnh đề cho mỗi ô:

$$x(1,1,1) \vee x(1,1,2) \vee x(1,1,3) \vee x(1,1,4)$$

Mỗi ô chứa nhiều nhất một giá trị — một mệnh đề "không thể cùng đúng" cho mỗi cặp giá trị:

$\neg x(1,1,1) \vee \neg x(1,1,2)$ $\neg x(1,1,1) \vee \neg x(1,1,3)$ … và tiếp tục như vậy cho cả sáu cặp.

Mỗi hàng chứa mọi giá trị — với hàng 1 và giá trị 3: ít nhất một lần,

$$x(1,1,3) \vee x(1,2,3) \vee x(1,3,3) \vee x(1,4,3)$$

và nhiều nhất một lần: $\neg x(1,1,3) \vee \neg x(1,2,3)$, rồi tương tự cho mỗi cặp ô trong hàng.
Cột và khối — các nhóm mệnh đề giống hệt; chỉ thay nhóm ô. Với khối trên cùng bên trái và giá trị 2:

$$x(1,1,2) \vee x(1,2,2) \vee x(2,1,2) \vee x(2,2,2)$$

cộng với các mệnh đề "không thể cùng đúng" cho từng cặp.
Các số gợi ý đã in — phần đơn giản nhất: mỗi gợi ý là một mệnh đề chỉ có một biến. Một số 3 được in ở góc trên bên trái trở thành mệnh đề

$$x(1,1,3)$$

Phép AND của tất cả những điều này khả thỏa mãn đúng khi Sudoku có lời giải — và một phép gán thỏa mãn *chính là* lời giải: đọc xem những x(r,c,v) nào là đúng rồi điền vào lưới. Với Sudoku 9×9, ta có 729 biến và vài nghìn mệnh đề, thứ mà một bộ giải SAT hiện đại xử lý trong vài mili giây. Hãy chú ý mệnh đề gợi ý x(1,1,3): nó nói "ô này bằng chính xác 3", không phải "các ô này đều khác nhau" — chính sự bất đối xứng đó sẽ buộc ta dùng thêm một mẹo cho các ô gợi ý trong ghi chú về giao thức ở phía dưới.
Từ SAT sang Sudoku. Bài báo cần chiều ngược lại, khó hơn: với một công thức SAT *bất kỳ*, hãy xây một mega-Sudoku có lời giải đúng khi công thức đó có lời giải. Các quy tắc tự nhiên của Sudoku chỉ có thể nói "các ô này đều khác nhau", nên các ràng buộc logic tùy ý phải được *xây dựng* — và đó chính là vai trò của các gadget. Một gadget là một cụm ô nhỏ được chế tạo sẵn, một cụm cho mỗi mệnh đề của công thức, trong đó các ô được chỉ định đóng vai trò biến (ký hiệu chúng chứa mã hóa đúng hoặc sai), còn các ràng buộc bên trong cụm được thiết kế sao cho những cách điền hợp lệ duy nhất tương ứng với các phép gán thỏa mãn mệnh đề đó. Đây là kỹ thuật tiêu chuẩn trong các chứng minh NP-đầy đủ; với Sudoku tổng quát, Yato và Seta đã xây dựng nó vào năm 2003.
Gộp hai chiều lại, Sudoku N×N và SAT là cùng một bài toán khoác hai bộ trang phục khác nhau. Đó là điều cho phép bài viết này — và bài báo gốc — kể câu chuyện về toàn bộ NP bằng lưới và ký hiệu.

Witness vẫn rất dễ hình dung. Alice biết một cách điền MegaSudoku hoàn chỉnh và hợp lệ. Bob muốn tin rằng cách điền như vậy tồn tại, nhưng Alice không muốn tiết lộ nó. Nếu cô gửi toàn bộ lưới, Bob sẽ tin, nhưng bí mật biến mất.

Trong phiên bản zero-knowledge cổ điển, Alice và Bob tương tác. Một mô hình tinh thần kiểu cũ dùng các quân được úp kín. Alice che lưới đã giải, bí mật đổi tên các ký hiệu trước mỗi vòng, rồi cho Bob kiểm tra một ràng buộc cục bộ được chọn ngẫu nhiên: một hàng, một cột, một khối hoặc một gadget. Nếu các ô được mở cho thấy những ký hiệu đều khác nhau, Bob có thêm niềm tin. Sau đó mọi thứ lại được che kín và các ký hiệu được đổi tên mới. (Có một chi tiết: các gợi ý cho sẵn của câu đố cần thêm một mẹo, vì đổi tên ký hiệu cũng che luôn chúng. Ghi chú bên dưới giải thích cách các giao thức cổ điển xử lý chuyện này; hình dung đơn giản ở đây là đủ cho phần sau.)

Các giao thức cổ điển thực sự xử lý ô gợi ý như thế nào

Mẹo đổi tên có một điểm mù. Các quy tắc hàng, cột và khối đều nói "các ô này đều khác nhau", và tính chất *đều khác nhau* vẫn giữ nguyên dưới mọi phép đổi tên ký hiệu. Nhưng một gợi ý nói "ô này chứa chính xác 5", và sau khi đổi tên Bob chỉ thấy $\sigma(5)$ — một ký hiệu đã bị che — mà không biết phép đổi tên σ. Anh ta không thể kiểm tra gì cả. Nếu không sửa điểm này, Alice có thể chứng minh rằng *một* lưới hợp lệ nào đó tồn tại trong khi bỏ qua hoàn toàn các gợi ý đã in, điều không chứng minh gì về *câu đố này*. Tài liệu cổ điển có hai cách sửa tiêu chuẩn.

Bảng màu. Thêm một hàng phụ gồm N ô vào lưới ẩn — một bảng màu mà Alice điền các ký hiệu 1…N theo thứ tự công khai cố định, rồi đổi tên nó cùng với mọi thứ khác, để nó chứa $\sigma(1)$…$\sigma(N)$. Thử thách ngẫu nhiên của Bob giờ có thêm một lựa chọn. Ngoài việc chọn một hàng, cột, khối hoặc gadget để mở, anh ta có thể chọn **bảng màu cộng với một ô gợi ý**. Alice mở cả hai; bảng màu cho biết phép đổi tên của vòng đó, và Bob kiểm tra rằng ô gợi ý hiển thị đúng phiên bản đã đổi tên của gợi ý được in. Việc này vẫn là zero-knowledge vì Bob chỉ biết σ — được lấy mới ở mỗi vòng và tự nó vô dụng — cùng giá trị của một ô mà anh ta vốn đã biết từ câu đố. Không có gì về các ô bí mật bị lộ, và một bộ mô phỏng có thể giả lập cảnh này bằng cách chọn một σ ngẫu nhiên. Giao thức có tính đúng đắn vì Alice gian lận sẽ bị bắt với một xác suất cố định ở mỗi vòng, và các vòng được lặp lại cho đến khi xác suất nghi ngờ trở nên không đáng kể.

Biên dịch các gợi ý đi. Một biến thể có cấu trúc hơn loại bỏ thử thách đặc biệt thay vì thêm nó. Thay vì *xác minh* giá trị gợi ý, hãy ép nó bằng các ràng buộc khác biệt: nối ô gợi ý với mọi ô của bảng màu trừ ô mang chính giá trị của nó — "khác $\sigma(1)$, khác $\sigma(2)$, …, khác mọi thứ ngoại trừ $\sigma(5)$." Ký hiệu duy nhất ô đó có thể hợp lệ chứa là giá trị gợi ý. Giờ mọi ràng buộc lại có dạng "hai ô này khác nhau" — bất biến dưới phép đổi tên, có thể kiểm tra hệt như một hàng. Đây cũng là thủ pháp dùng cho các đỉnh đã được tô màu sẵn trong giao thức tô màu đồ thị cổ điển, và đúng với tinh thần của từ *gadget* ở trên: trong cách nhìn MegaSudoku-như-SAT, các gợi ý được biên dịch thành gadget bất đẳng thức giống mọi ràng buộc khác.

Giao thức vật lý. Giao thức dùng thẻ ngoài đời cho Sudoku (Gradwohl, Naor, Pinkas và Rothblum, 2007) hoàn toàn không dùng đổi tên và xử lý gợi ý trước cả khi việc che giấu bắt đầu. Với mỗi ô, Alice đặt ba thẻ giống nhau mang giá trị của ô — úp xuống với ô bí

> mật, nhưng **ngửa lên với ô gợi ý**, để Bob tận mắt thấy các gợi ý được tôn trọng trước khi các thẻ được lật úp. Sau đó một thẻ từ mỗi ô được đưa vào gói của hàng, một thẻ vào gói của cột, một thẻ vào gói của khối; mỗi gói được xáo rồi mở ra, và Bob kiểm tra nó chứa đủ N ký hiệu. Việc xáo làm mất thông tin vị trí (đó là phần zero-knowledge), còn các gợi ý đã được cố định ngay từ lúc chia thẻ.
> Dù dùng cách nào, bài học vẫn là điều bài viết này liên tục quay lại: một giao thức zero-knowledge là phép ghi sổ rất cẩn thận về *những* sự thật nào còn tồn tại sau khi che giấu. Đổi tên bảo toàn "tất cả khác nhau" và xóa mất "bằng 5" — vì thế "bằng 5" phải được đưa trở lại bằng cách khác.

Đó không phải giao thức trong bài báo. Nó chỉ là mô hình tinh thần cho zero-knowledge cổ điển:

- Alice và Bob trao đổi qua lại.
- Bob chọn các phép kiểm tra ngẫu nhiên.
- Alice chỉ tiết lộ tính nhất quán cục bộ, không phải toàn bộ lời giải.
- Chứng minh về tính riêng tư hoạt động bằng cách cho thấy góc nhìn của Bob có thể được tạo ra mà không cần lời giải bí mật của Alice.

Vì vậy, zero-knowledge cổ điển được xây quanh một sự thật khẳng định:

> Một bộ mô phỏng thực sự tồn tại.

Bây giờ hãy bỏ các phần dễ chịu đi. Alice gửi một chuỗi bằng chứng duy nhất rồi rời đi. Không có thiết lập đáng tin cậy, không có chuỗi ngẫu nhiên chung được chuẩn bị trước, và Bob tuyệt đối không được chấp nhận một câu đố sai. Đây là bối cảnh mà zero-knowledge cổ điển không thể tồn tại.

Ta cần thêm một nhân vật nữa trước khi đến mẹo chính. Hãy cố định một **bộ quy tắc** (*rulebook*): một hệ chứng minh hình thức theo nghĩa của nhà logic học — một tập tiên đề cố định cộng với các quy tắc cơ học để kiểm tra các bằng chứng toán học được viết ra. ZFC, hệ tiên đề tiêu chuẩn của toán học, là ví dụ kinh điển. Từ đây trở đi, mọi phát biểu đều tương đối so với một bộ quy tắc được chọn trước, và lựa chọn này linh hoạt: cấu trúc hoạt động với bất kỳ bộ quy tắc nào bạn cố định, kể cả ZFC.

(Một lưu ý về từ ngữ, mượn chính từ bài báo: "proof system" ở đây luôn có nghĩa là **hệ chứng minh hình thức**, tức bộ quy tắc dùng để kiểm tra các bằng chứng toán học — tuyệt đối không phải các thông điệp Alice gửi. Bộ máy của Alice và Bob được gọi là "người chứng minh và người xác minh".)

Phiên bản kiểu Gödel giữ câu chuyện MegaSudoku nhưng thay đổi bằng chứng.

Chọn một hệ ràng buộc thứ hai có cùng kích thước hiển thị, gọi nó là **D**. Trong câu chuyện, S và D là hai câu đố MegaSudoku(n) cùng định dạng. Phía sau hậu trường, D có thể bắt đầu từ một công thức logic khó với kích thước khác; nếu cần, có thể đệm thêm các ràng buộc giả vô

hại để nó vừa cùng một lưới. D được xây từ một công thức logic thực sự **không khả thỏa mãn** (*unsatisfiable*): không có cách gán giá trị nào khiến mọi ràng buộc của nó cùng đúng, giống như một câu đố hỏng không có lưới hoàn chỉnh hợp lệ. Một ví dụ đồ chơi là công thức đòi hỏi đồng thời "X đúng" và "X sai". Vì vậy D không có cách điền hợp lệ.

Nhưng D không được là một câu đố hỏng *dễ bị lật tẩy*. Ví dụ đồ chơi ở trên thất bại ở điểm này: bất kỳ bộ quy tắc nào cũng bác bỏ "X và không-X" chỉ trong một dòng. D phải sai theo cách mà bộ quy tắc đã chọn không thể chứng nhận bằng một lập luận ngắn. Nếu bộ quy tắc có thể bác bỏ D bằng một bằng chứng ngắn, câu chuyện phía dưới sẽ sụp đổ: con đường thay thế có thể tạo ra bằng chứng mà không cần bí mật của Alice sẽ bị loại trừ một cách hình thức, kéo theo bảo đảm riêng tư. Vì vậy D được chọn từ một họ mà bộ quy tắc cố định không thể bác bỏ hiệu quả: bên trong bộ quy tắc đó không có bằng chứng ngắn rằng D không có lời giải.

Bằng chứng một thông điệp của Alice khi ấy nói về một mệnh đề "hoặc":

> hoặc MegaSudoku thật S có lời giải, hoặc mồi nhử D có lời giải.

Đây là mắt xích logic. D **không** được tạo ra bằng cách kỳ diệu nào khiến S trở thành đúng. Bằng chứng không lập luận "D không có lời giải, do đó S có lời giải". Nó chứng minh phép tuyển **S hoặc D**. Tính đúng đắn hoàn hảo nói rằng một phép tuyển sai không thể có bằng chứng hợp lệ. Vì trên thực tế D là sai — nó không có lời giải — cách duy nhất để phép tuyển đúng là S phải đúng. Vì vậy nếu bằng chứng được chấp nhận, S phải có lời giải. Mồi nhử không thể biến một S sai thành đúng.

Nhưng với phần mang phong cách zero-knowledge, hãy hỏi điều gì sẽ xảy ra nếu D *thực sự* có lời giải. Lời giải mồi nhử đó sẽ đóng vai trò một witness thay thế. Nó cho phép ai đó tạo bằng chứng mà không biết lời giải MegaSudoku thật của Alice — nói cách khác, nó tạo ra một bộ mô phỏng. Trong thực tế D không có lời giải, nên con đường mô phỏng này đóng. Mấu chốt là bộ quy tắc không thể chứng minh hiệu quả rằng nó đã đóng.

Vì vậy D có hai nhiệm vụ. Với **tính đúng đắn**, D là sai, nên một bằng chứng hợp lệ của "S hoặc D" buộc S phải đúng. Với **zero-knowledge hiệu dụng**, D khó bị bác bỏ, nên bộ quy tắc không thể nhanh chóng loại trừ con đường mồi nhử vốn sẽ làm cho việc mô phỏng trở nên khả thi.

Vì thế phép thử an ninh không còn là:

> Ta có thể chứng minh rằng một bộ mô phỏng thực sự tồn tại hay không?

Nó trở thành:

> Bộ quy tắc của bạn có thể chứng minh hiệu quả rằng bộ mô phỏng là bất khả thi hay không?

Nếu câu trả lời là không, một điều mạnh đến bất ngờ xảy ra: mọi bảo đảm an ninh mà (a) có thể quan sát bằng cách chạy một phép thử, và (b) có thể chứng minh — bên trong bộ quy tắc

đó — là hệ quả của sự tồn tại của bộ mô phỏng, đều thực sự được giữ. Một cuộc tấn công thành công vào bất kỳ bảo đảm nào trong số đó tự nó sẽ tạo thành phép bác bỏ ngắn còn thiếu, nhưng phép bác bỏ ngắn đó không tồn tại. Đó chính là phần "hiệu dụng" trong zero-knowledge hiệu dụng.

Vì vậy, đối chiếu trong lớp học là:

Zero-knowledge cổ điển: bằng chứng an toàn vì một bộ mô phỏng tồn tại.

Zero-knowledge hiệu dụng kiểu Gödel: bằng chứng được coi là an toàn đối với các phép thử an ninh quan sát được vì bộ quy tắc không thể chứng minh hiệu quả rằng bộ mô phỏng là bất khả thi.

Khẳng định thứ hai yếu hơn. Và chính vì yếu hơn nên bài báo có thể giữ ba đặc điểm đã phá vỡ phiên bản cổ điển: một thông điệp, không thiết lập và tính đúng đắn hoàn hảo.

Phép thử mới: bạn không thể chứng minh bộ mô phỏng vắng mặt

Sự nới lỏng của Ilango thay đổi câu hỏi.

Zero-knowledge cổ điển hỏi:

> Một bộ mô phỏng có tồn tại không?

Zero-knowledge hiệu dụng hỏi một điều yếu hơn:

> Bộ quy tắc bạn chọn có thể chứng minh hiệu quả rằng không có bộ mô phỏng nào tồn tại không?

Điều này nghe như một cách lách kỹ thuật, nhưng nó là ý tưởng cốt lõi. Cấu trúc sống trong một trạng thái kỳ lạ: trên thực tế bộ mô phỏng không tồn tại — bài báo nói rõ điều này — nhưng bộ quy tắc bạn cố định không thể chứng minh hiệu quả rằng nó không tồn tại. Nếu mọi hậu quả xấu mà bạn quan tâm đều đòi hỏi một phép bác bỏ như vậy, hệ thống vẫn hành xử như zero-knowledge đối với các hậu quả đó.

Đây là nơi Gödel xuất hiện. Không phải để trang trí, và không phải theo nghĩa "Gödel làm mật mã an toàn". Mối liên hệ nằm ở lý thuyết chứng minh. Một bộ quy tắc được gọi là **tối ưu** nếu, theo một nghĩa chính xác, nó là bộ tốt nhất có thể: bất cứ khi nào một bộ quy tắc nào đó có thể bác bỏ một công thức thuộc loại liên quan bằng một bằng chứng ngắn, bộ quy tắc tối ưu cũng có thể làm vậy với bằng chứng dài hơn nhiều nhất theo một hệ số đa thức. Krajíček và Pudlák đưa ra giả thuyết năm 1989 rằng **không tồn tại hệ chứng minh tối ưu**: dù bạn cố định bộ quy tắc nào, vẫn có một bộ quy tắc khác chứng minh một họ mệnh đề đúng nào đó ngắn gọn

hơn rất nhiều. Đây là một trong những giả thuyết mở trung tâm của độ phức tạp chứng minh, và là họ hàng hữu hạn, theo lý thuyết độ phức tạp, của định lý bất toàn Gödel: một số mệnh đề đúng không có bằng chứng ngắn trong bộ quy tắc bạn cố định — không phải vì về nguyên tắc chúng không thể chứng minh được, mà vì mỗi bộ quy tắc cố định đều bỏ sót một số sự thật ngắn mà không có bằng chứng ngắn.

Bài báo giả định giả thuyết này (ở dạng "vô hạn lần" hơi mạnh hơn, dạng thường dùng khi các giả thuyết được đưa vào mật mã học). Phần thưởng, nhờ một định lý của Krajíček và Pudlák, rất cụ thể: với mỗi bộ quy tắc tồn tại một dãy công thức thực sự không khả thỏa mãn nhưng bộ quy tắc không thể bác bỏ bằng các bằng chứng ngắn — và quan trọng là một thuật toán hiệu quả có thể *tạo* chúng. Tính chất cuối này, tính đồng nhất (*uniformity*), biến toàn bộ ý tưởng từ một phát biểu tồn tại thành một thuật toán Alice thực sự có thể chạy: các mồi nhử D của cô đi ra từ một dây chuyền, không phải xuất hiện từ hư không.

Nước đi mật mã học là đưa sự thiếu hụt năng lực chứng minh đó vào sử dụng.

Cấu trúc đang làm gì

Đây là cấu trúc của bài báo, bỏ bớt chi tiết để chỉ giữ hình dạng.

Hãy cố định một bộ quy tắc — ví dụ ZFC. Theo giả định về độ phức tạp chứng minh, tồn tại một dãy công thức có thể tạo hiệu quả, thực sự không khả thỏa mãn, nhưng bộ quy tắc không có bằng chứng ngắn rằng chúng không khả thỏa mãn.

Bây giờ xây một bằng chứng một thông điệp có dạng:

> hoặc mệnh đề thật là khả thỏa mãn, hoặc công thức khó đặc biệt này là khả thỏa mãn.

Công thức khó đặc biệt thực sự không khả thỏa mãn. Vì vậy nếu bộ máy chứng minh nền có tính đúng đắn hoàn hảo, việc chấp nhận thông điệp vẫn có nghĩa là mệnh đề thật đúng. Đó là tính đúng đắn hoàn hảo.

Nhưng với an ninh kiểu zero-knowledge, hãy tưởng tượng công thức khó đặc biệt *có thể* được thỏa mãn. Khi ấy witness của nó có thể được dùng để mô phỏng các bằng chứng mà không cần biết witness thật. Trong thực tế công thức đó không khả thỏa mãn — nhưng bộ quy tắc không thể chứng minh hiệu quả điều đó. Vì vậy nó không thể chứng minh hiệu quả rằng bộ mô phỏng là bất khả thi.

Đó là bản lề của toàn bộ ý tưởng. Hệ thống không che bí mật bằng cách tạo ra một bộ mô phỏng cổ điển. Nó che bí mật, đối với một lớp lớn các phép thử an ninh quan sát được, phía sau việc bộ quy tắc không thể chứng nhận rằng bộ mô phỏng vắng mặt.

Bài báo tuyên bố điều gì

Định lý chính có nhiều lớp. Kết quả cốt lõi là thế này:

Dưới một giả định mật mã học tiêu chuẩn — sự tồn tại của **bằng chứng không tương tác có tính không phân biệt nhân chứng** (*non-interactive witness indistinguishable proofs*), một đối tượng được nghiên cứu kỹ và suy ra từ nhiều gói giả định đã được thiết lập — cùng giả thuyết của độ phức tạp chứng minh rằng **không tồn tại hệ chứng minh tối ưu (theo dạng "vô hạn lần")**, bài báo xây dựng, với mọi lựa chọn bộ quy tắc, một người chứng minh và người xác minh một thông điệp cho NP/SAT, có **tính đúng đắn hoàn hảo**, không thiết lập, và zero-knowledge hiệu dụng tương đối với bộ quy tắc đó. (NP/SAT là "mẫu số chung khó nhất" tiêu chuẩn của những bài toán giống câu đố; mega-Sudoku là một bộ trang phục của nó.)

Với tuyên bố rộng hơn về việc bảo toàn các thuộc tính an ninh có thể kiểm nghiệm để bác bỏ, bài báo thêm một giả định tiêu chuẩn nữa, niềm tin khử ngẫu nhiên **P** = **BPP** (nói gần đúng: tính ngẫu nhiên không cho thuật toán thêm sức mạnh thiết yếu).

Dịch khỏi ngôn ngữ định lý:

- Bằng chứng chỉ là một thông điệp.
- Không có thiết lập đáng tin cậy.
- Các mệnh đề sai không thể được chứng minh.
- Người chứng minh không phải zero-knowledge cổ điển — nó không có bộ mô phỏng.
- Nhưng từng hệ quả an ninh dựa trên trò chơi và có thể kiểm nghiệm để bác bỏ của zero-knowledge cổ điển đều có thể đạt được trong bối cảnh này.

"Có thể kiểm nghiệm để bác bỏ" (*falsifiable*) là từ quan trọng. Nó có nghĩa là một thất bại an ninh có thể được kiểm tra bằng cách cho đối thủ chạy trong một trò chơi. Nhiều định nghĩa an ninh mật mã có dạng này: đối thủ có thể phân biệt hai bản mã, đảo ngược một hàm, khôi phục witness hay thắng một thí nghiệm quy định sẵn không? Định lý cho một người chứng minh cho từng thuộc tính như vậy, mỗi lần một thuộc tính. Một người chứng minh duy nhất có *mọi* thuộc tính có thể kiểm nghiệm để bác bỏ cùng lúc có lẽ là bất khả thi — cuộc tấn công tái sử dụng cũ ("Bob có thể đưa bằng chứng cho người khác") tự nó cũng là một thuộc tính có thể kiểm nghiệm để bác bỏ, và nó thực sự thất bại ở đây. Đề xuất của bài báo là một người chứng minh duy nhất có thể hợp lý bao phủ mọi thuộc tính có thể kiểm nghiệm để bác bỏ *tự nhiên* — những thuộc tính thực sự xuất hiện trong thực hành mật mã học — nhưng phần đó là một định lý có điều kiện dựa trên một khái niệm "tự nhiên" không hoàn toàn hình thức, cộng thêm một giả thuyết được nêu rõ. Bảo đảm nhắm vào những thất bại quan sát được, không phải mọi ý nghĩa triết học hay dựa trên mô phỏng của tính bí mật.

Một hệ quả cụ thể đáng gọi tên: cấu trúc tạo ra các bằng chứng **che giấu nhân chứng** (*witness hiding*) không tương tác đầu tiên với người chứng minh đồng nhất — "một bằng chứng rằng câu đố có lời giải không giúp bạn tìm ra lời giải", không tương tác và không thiết lập — một đối tượng nghe khiêm tốn nhưng đã chống lại việc xây dựng trong nhiều thập niên.

Điều này không nói lên điều gì

Đây là phần giữ cho bài viết trung thực.

Nó **không** nói rằng các định lý bất khả thi cũ là sai. Cấu trúc tránh chúng bằng cách thay đổi định nghĩa.

Nó **không** cho zero-knowledge cổ điển thông thường với không tương tác, không thiết lập và tính đúng đắn hoàn hảo. Bài báo nói rõ người chứng minh được xây dựng không có bộ mô phỏng.

Nó **không** có nghĩa bằng chứng không thể được tái sử dụng. Một bằng chứng một thông điệp vẫn có thể được đưa cho người khác; bài báo không bảo toàn các thuộc tính kiểu chối bỏ (*deniability*). (Zero-knowledge không tương tác với thiết lập đáng tin cậy cũng có giới hạn này.)

Nó **không** có nghĩa đây là một giao thức thực dụng sẵn sàng triển khai. Đây là lý thuyết độ phức tạp và nền tảng mật mã học. Kết quả phụ thuộc vào các giả định lớn từ độ phức tạp chứng minh và mật mã học, và cấu trúc nói về điều gì có thể làm được về nguyên tắc.

Nó **không** biến "Gödel" thành một cơ chế an ninh thần kỳ. Mối liên hệ với Gödel đi qua các hệ chứng minh, hệ chứng minh tối ưu và những tương tự hữu hạn của bất toàn. Trực giác hữu dụng không phải "bất toàn bảo vệ mật khẩu của bạn". Mà là: nếu một bộ quy tắc không thể chứng minh hiệu quả rằng một bộ mô phỏng là bất khả thi, thì những cuộc tấn công đòi hỏi bằng chứng đó có thể bị chặn ở cấp độ định nghĩa an ninh.

Vì sao nó vẫn thú vị

Mật mã học thường biến độ khó thành an toàn. Phân tích thừa số khó, nên các giả định kiểu RSA trở nên hữu ích. Các bài toán mạng tinh thể khó, nên mật mã mạng tinh thể trở nên hữu ích. Ở đây độ khó kỳ lạ hơn: không phải "khó tính ra một bí mật", mà là "khó chứng minh rằng một đối tượng chứng minh nhất định không thể tồn tại".

Đó là lý do bài báo có cảm giác khác thường. Nó đối xử với tiên đề và bộ quy tắc gần như các tài nguyên mật mã. Kết quả bất khả thi thông thường nói rằng có một căng thẳng giữa tính đúng đắn và mô phỏng. Nước đi của Ilango là đặt căng thẳng đó sau một bức màn lý thuyết chứng minh: bộ mô phỏng vắng mặt, nhưng hệ hình thức không thể hiệu quả phơi bày sự vắng mặt đó.

Với người đọc, điều đáng ngạc nhiên không phải là hệ này sẽ thay thế các hệ zero-knowledge hiện nay. Có lẽ nó sẽ không, ít nhất không trực tiếp. Điều đáng ngạc nhiên là một giới hạn từ logic toán học có thể được dùng theo hướng xây dựng: không chỉ như một bức tường, mà như một lớp che phủ.

Bằng chứng mạnh đến đâu?

Đây là một bài báo về định lý, nên "bằng chứng" mang nghĩa khác với trong sinh học hay thiên văn học. Câu hỏi không phải liệu một thí nghiệm đã được lặp lại hay chưa. Câu hỏi là liệu các định nghĩa, giả định và chuỗi chứng minh có hỗ trợ tuyên bố hay không.

Chứng minh là hình thức, và bài báo nêu rõ các giả định. Các giả định không hề tùy tiện. Bằng chứng không tương tác có tính không phân biệt nhân chứng là đối tượng tiêu chuẩn trong mật mã học và suy ra từ nhiều gói giả định đã được thiết lập. Giả thuyết không-hệ-chứng-minh-tối-ưu là một giả thuyết trung tâm trong độ phức tạp chứng minh. **P** = **BPP** là một niềm tin khử ngẫu nhiên tiêu chuẩn, chỉ dùng cho định lý rộng hơn về các thuộc tính có thể kiểm nghiệm để bác bỏ.

Bài báo còn lập luận rằng những giả định này là cái giá phù hợp, không phải một giàn giáo tùy ý: nó chứng minh một chiều ngược cho thấy chúng về cơ bản là cần thiết — nếu các cấu trúc như thế này tồn tại, thì bằng chứng không tương tác có tính không phân biệt nhân chứng phải tồn tại, và (nếu chấp nhận các hàm một chiều tiêu chuẩn) không thể có hệ chứng minh tối ưu. Các giả định cũng mang tính "đôi bên cùng thắng": bác bỏ bất kỳ giả định nào trong số đó tự nó sẽ là một phát hiện mang tính cột mốc trong độ phức tạp chứng minh, mật mã học hoặc lý thuyết độ phức tạp.

Nhưng vì kết quả có điều kiện, mức tin cậy cũng có điều kiện. Nếu các giả định đó sai, cách diễn giải định lý thay đổi. Và ngay cả khi chúng đúng, bảo đảm vẫn không phải zero-knowledge cổ điển đầy đủ; nó là phiên bản nới lỏng, dựa trên lý thuyết chứng minh của bài báo.

Vì vậy mức tin cậy phù hợp là cao rằng bài báo thiết lập một kết quả khả thi có điều kiện và nhất quán; trung bình rằng các giả định của nó mô tả thế giới mật mã mà chúng ta thực sự sống trong đó; và thấp đối với bất kỳ hệ quả thực tiễn tức thời nào.

Vì sao nó quan trọng

Bài báo mở ra một con đường vốn được cho là đã đóng.

Lý thuyết cổ điển nói: zero-knowledge đầy đủ không thể là một thông điệp nếu không có thiết lập, và không thể có tính đúng đắn hoàn hảo. Bài báo của Ilango nói: nếu ta chỉ yêu cầu những hệ quả của zero-knowledge có thể được kiểm tra trong các trò chơi an ninh, và nếu ta cho phép định nghĩa an ninh phụ thuộc vào điều một bộ quy tắc có thể hay không thể bác bỏ hiệu quả, thì nhiều hành vi hữu ích có thể được khôi phục — với một thông điệp, không thiết lập và tính đúng đắn hoàn hảo.

Đó không phải một chỉnh sửa định nghĩa nhỏ. Nó là một cách khác để nghĩ về các bảo đảm mật mã. Thay vì chỉ hỏi thứ gì tồn tại, hãy hỏi bộ quy tắc của bạn có thể loại trừ điều gì. Thay vì coi tính không chứng minh được như một phiền toái triết học, hãy dùng nó như cấu trúc.

Thế giới thực dụng có thể không thay đổi vào ngày mai. Nhưng bản đồ khái niệm thì thay đổi.

Giờ tồn tại một nghĩa hình thức trong đó "không ai có thể chứng minh hiệu quả rằng bí mật đã bị lộ" có thể đủ mạnh để khôi phục nhiều bảo vệ dựa trên trò chơi mà ta muốn từ "bí mật không bị lộ".

Đó là lý do Gödel xuất hiện trong tiêu đề.

Tóm tắt gọn

Bằng chứng zero-knowledge cho phép một người chứng minh thuyết phục người xác minh rằng một mệnh đề là đúng mà không tiết lộ witness. Các kết quả bất khả thi cổ điển nói rằng zero-knowledge không thể bị ép vào một thông điệp duy nhất nếu không có thiết lập, và không thể có tính đúng đắn hoàn hảo. Bài báo của Rahul Ilango không bác bỏ các bất khả thi đó. Nó định nghĩa một khái niệm yếu hơn, **zero-knowledge hiệu dụng**: thay vì yêu cầu một bộ mô phỏng thực sự tồn tại, nó yêu cầu rằng một hệ chứng minh được chọn — một bộ quy tắc hình thức như ZFC — không thể chứng minh hiệu quả rằng không có bộ mô phỏng. Dưới các giả định lớn từ mật mã học (bằng chứng không tương tác có tính không phân biệt nhân chứng) và độ phức tạp chứng minh (không tồn tại hệ chứng minh tối ưu), bài báo xây các người chứng minh một thông điệp cho NP/SAT, không thiết lập và có tính đúng đắn hoàn hảo, đạt từng hệ quả an ninh dựa trên trò chơi, có thể kiểm nghiệm để bác bỏ của zero-knowledge. Một người chứng minh duy nhất bao phủ mọi thuộc tính "tự nhiên" như vậy là một mở rộng xa hơn, một phần còn mang tính giả thuyết — và bao phủ theo nghĩa đen *mọi* thuộc tính có thể kiểm nghiệm để bác bỏ có lẽ là bất khả thi, vì bằng chứng vẫn có thể tái sử dụng. Kết quả là lý thuyết và có điều kiện, không phải một primitive mật mã đã triển khai, nhưng nó cho thấy một cách mới để dùng tính không chứng minh được trong lý thuyết chứng minh như một tài nguyên mật mã.

Kiểm tra không cường điệu

Bài báo cho thấy gì: Dưới các giả định được nêu, có thể xây các người chứng minh một thông điệp, không thiết lập, có tính đúng đắn hoàn hảo cho NP/SAT, zero-knowledge hiệu dụng tương đối với bất kỳ hệ chứng minh được chọn nào, và đạt từng hệ quả an ninh dựa trên trò chơi, có thể kiểm nghiệm để bác bỏ của zero-knowledge cổ điển.

Điều gì hợp lý nhưng chưa được chứng minh vô điều kiện: Các giả định cần thiết từ độ phức tạp chứng minh và mật mã học thực sự đúng. Chúng là các giả định nghiêm túc, được nghiên cứu kỹ — và bài báo cho thấy chúng về cơ bản vừa cần vừa đủ — nhưng vẫn là giả định.

Điều bài báo không cho thấy: Zero-knowledge cổ điển với không tương tác, không thiết lập và tính đúng đắn hoàn hảo; một hệ thống thực dụng sẵn sàng triển khai; khả năng chối bỏ hay không tái sử dụng bằng chứng; hoặc định lý bất toàn Gödel tự nó làm mật mã an toàn.

Giới hạn chính: Bảo đảm là một sự nới lỏng của zero-knowledge; phiên bản rộng nhất phụ thuộc vào nhiều giả định; các tuyên bố về một người chứng minh phổ quát duy nhất vẫn một phần mang tính giả thuyết; và kết quả chủ yếu thuộc nền tảng lý thuyết.

Một độc giả phổ thông nên tin ở mức nào? Cao rằng đây là một kết quả lý thuyết có điều kiện quan trọng nếu chấp nhận các định nghĩa. Trung bình rằng các giả định phản ánh thực tế. Thấp với triển khai thực tiễn ngay lập tức. Điều nên nhớ là: bài báo không phá vỡ các kết quả bất khả thi của zero-knowledge; nó tìm một con đường mới dựa trên lý thuyết chứng minh để đi vòng qua những phần của chúng quan trọng đối với nhiều trò chơi an ninh.

Nguồn

Ilango, IACR ePrint 2025/1296
https://eprint.iacr.org/2025/1296

FOCS 2025, DOI 10.1109/FOCS63196.2025.00058
https://doi.org/10.1109/FOCS63196.2025.00058