Cryptanalysis of a knot-based key exchange

Simon-Philipp Merz
Applied Cryptography Group, ETH Zurich

Abstract. We present an efficient attack on a knot-based Diffie–Hellman key exchange proposed by Sconza and Wildi. In the proposal, the two parties exchange oriented knots, combine them under connected sum to obtain a common knot, and derive the shared secret by evaluating a finite type invariant of degree m on it. We show the scheme is insecure for *every* choice of finite type invariant. The shared secret can be computed from the public transcript at roughly three times the cost of running the scheme honestly. The attack maps knots into a truncated Polyak space $\mathcal{P}_m$, in which connected sum becomes multiplication and every element of the image is invertible, so the public knot can simply be divided out. This bypasses all countermeasures put in place by the proposed protocol.

Independently of the attack, we show that the key space is too small for the parameters proposed. A degree-m invariant takes $O(c^m)$ values on knots represented with diagrams consisting of c crossings. For the suggested crossing number, reaching the 128 bits claimed would require $m \geq 10$, at which point a single evaluation of the invariant costs in the order of 2^{50} operations.

1. Introduction

An oriented knot in S^3 is a closed, non-self-intersecting loop of string placed in a 3-dimensional sphere that has a chosen direction. In this paper we will simply refer to these objects as knots. The connected sum $(K_1 \# K_2)$ of two knots K_1, K_2 is a single composite knot formed by cutting a small arc from each of the two separate knots and joining their open ends together without adding extra crossings. The set of knots becomes a commutative monoid under the operation of connected sum, with the unknot as the identity. Moreover, every knot factors uniquely into simpler, indecomposable prime knots.

The idea of building public-key cryptography from knots recurs in the literature, for instance in the key exchange of Sconza and Wildi [7] that we cryptanalyse here, and it is also raised in public fora such as MathOverflow[1] and Reddit[2]. The appeal is easy to see as combining oriented knots via a connected sum is cheap, and factoring knots into simpler, prime building blocks appears to be a computationally difficult problem.

We would like to state a structural concern with using knots in cryptography from the start even though it is orthogonal to our attack later: Knots have no efficiently computable normal form. A party who must transmit a knot transmits a *diagram* representing it, and no canonical choice of diagram is known to be efficiently computable. The diagram actually sent therefore depends on how it was produced, and so potentially on the secret that produced it. A protocol whose security rests on a transmitted representation not leaking information is fragile by construction. While the construction in [7] recognises this as a problem and employs so-called "complexification" steps to obfuscate the secret within the representation of the knots, it remains open whether the measures are sufficient to stop leakage of secrets through their explicit representation.

Date: September 5, 2026.

[1] https://mathoverflow.net/questions/439286/knot-diffie-hellman

[2] https://www.reddit.com/r/cryptography/comments/1699e5t/knotbased_cryptography/

Moreover, even if a normal form existed there are potential failure modes. For the closely related braid groups it was shown that products of the form ABC presented in Garside normal form retain prefixes and suffixes of the normal forms of the secret factors, which can be enough to recover them [4]. However, the attack we will present in this paper does not exploit any concrete representation.

Sconza and Wildi propose to instantiate a Diffie–Hellman key exchange with knots acting on themselves by # as follows.

Protocol 1.1 ([7, Protocol 3.1])**.** The parties agree publicly on an alternating prime knot diagram K, a bound n, a finite type knot invariant V with values in $\mathbb{Z}$, and a hash function H.

(1) Alice picks a random alternating prime knot A with at most n crossings, forms $A \# K$, *complexifies* the resulting diagram, and sends it to Bob.
(2) Bob likewise sends a complexified diagram of $B \# K$.
(3) Alice forms the connected sum of her secret A with the diagram received from Bob, and Bob proceeds mutatis mutandis. The two (likely different) resulting diagrams present the same knot $A \# B \# K$, so evaluating V gives both parties the same integer, and the shared key is $H\big(V(A \# B \# K)\big)$.

In the proposed scheme the key is extracted using knot invariants, because the two parties hold usually different diagrams of the same knot $A \# B \# K$. Finite type invariants are chosen because they are efficiently computable and several specific choices for invariants are excluded because they are multiplicative under # which would trivially allow attackers to break the scheme (see [7, Sect. 4.3]).

Contributions. We show that the proposed scheme is insecure for *every* choice of finite type invariant. Instead of trying to divide out the knot K from the information observed on the wire during the proposed key exchange, we sidestep this problem. We show that any finite type invariant of degree m factors through a *homomorphism* φ_m where $\varphi_m(K)$ is invertible for *all* K. Using that φ_m is a homomorphism, we have

$$\varphi_m(A \# B \# K) = \varphi_m(A \# K)\, \varphi_m(K)^{-1}\, \varphi_m(B \# K),$$

where the shared key can be derived from the left hand side and the right hand side can be computed from public information for arbitrary choices of A, B, and K.

The whole attack costs roughly three evaluations of the invariant, about three times the cost of running the scheme honestly.

Additionally, we show that the key space of the proposed scheme is much smaller than assumed, for any choice of a single degree-m invariant to derive the shared key. At the crossing number proposed reaching 128 bits requires $m \geq 10$, at which point a single evaluation of the invariant, which the honest parties perform on every exchange, costs some 2^{50} operations.

Outline. Section 2 recalls the needed definitions and establishes the concatenation product. Section 3 develops the theory underlying the attack and gives the attack itself, and Section 4 discusses the low entropy in the proposed key space. We conclude in Section 5.

2. Preliminaries

We recall only what is needed to follow our attack, and for a more complete introduction to knots we refer to Roberts' "Knots Knotes" [6].

Definition 2.1. Let D be a knot diagram with c crossings. Choose a basepoint and traverse the diagram following the orientation of the knot, labelling the $2c$ passages through crossings by $1, \ldots, 2c$. The *Gauss diagram* of D consists of a circle with these $2c$ marked points, in which the two points belonging to a crossing are joined by an arrow directed from the over-passage to

the under-passage and decorated with the sign of the crossing. A *based arrow diagram* is the same object without the signs, i.e. a circle with a basepoint, finitely many marked points, and a matching of them into directed arrows, not necessarily realisable by a knot. Its *degree* is its number of arrows.

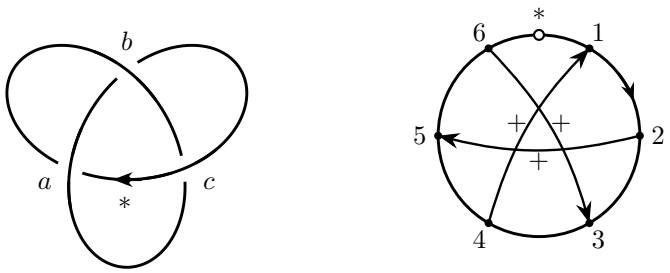

FIGURE 1. The right-handed trefoil knot and its Gauss diagram. The basepoint is marked $*$ and the arrow gives the direction of travel. Starting there one meets the crossings in the order a, b, c, a, b, c, which numbers the six passages $1, \dots, 6$. Each crossing contributes one arrow, directed from its over-passage to its under-passage and carrying its sign. The corresponding based arrow diagram is the same picture with the three signs erased.

An example using the simplest non-trivial knot is shown in Figure 1. Cutting the circle at the basepoint presents a based arrow diagram of degree k as a linearly ordered sequence of $2k$ endpoints with a perfect matching into arrows and a direction for each arrow. Note that there are finitely many such diagrams in each degree.

Definition 2.2. Let $\mathcal{D}$ be the $\mathbb{Q}$-vector space spanned by all based arrow diagrams. For a knot diagram D we write $D' \subseteq D$ for a *subdiagram*, obtained by deleting some arrows of the Gauss diagram of D and retaining the basepoint, and set

$$\varphi\colon \{\text{knot diagrams}\} \to \mathcal{D}, \qquad \varphi(D) = \sum_{D' \subseteq D} \varepsilon(D')\,[D'],$$

where $\varepsilon(D')$ is the product of the signs of the arrows of D', empty product 1, and $[D']$ is the based arrow diagram obtained from D' by forgetting the signs.

Equivalently, for a based arrow diagram α let $\langle \alpha, D \rangle$ denote the signed count of subdiagrams of D of pattern α, each contributing the product of the signs of its arrows, as in the Gauss diagram formulas of Polyak and Viro [5]. Then $\varphi(D) = \sum_\alpha \langle \alpha, D \rangle\, \alpha$, where the sum ranges over all based arrow diagrams, and the empty diagram occurs with coefficient 1.

Definition 2.3. The *Polyak space* $\mathcal{P}$ is the quotient of $\mathcal{D}$ by the relations given in [3, Sect. 2.5], and $\mathcal{P}_m$ is obtained by setting all diagrams of degree $> m$ to zero. We write $\pi\colon \mathcal{D} \to \mathcal{P}_m$ for the canonical projection, $\varphi_m = \pi \circ \varphi$ for the induced map on knot diagrams, which is a knot invariant, and $\mathbf{1} \in \mathcal{P}_m$ for the class of the empty diagram.

The relations are imposed precisely so that φ_m becomes invariant under the Reidemeister moves, i.e. the changes to the diagrams that relate any two equivalent knots.

Remark 2.4. Each relation is displayed as a picture in which only part of a diagram is drawn with the convention being that the undrawn part is the same in every term of the relation [3, Sect. 2.5]. In particular, this means each relation is a linear combination of based arrow diagrams that agree outside some arc J of the circle not containing the basepoint, and differ only inside J.

Definition 2.5. Let α, β be based arrow diagrams. Their *concatenation* $\alpha \cdot \beta$ is the based arrow diagram obtained by cutting both circles at their basepoints and juxtaposing the resulting sequences of endpoints, α first. Extending bilinearly makes $\mathcal{D}$ an associative $\mathbb{Q}$-algebra with unit the empty diagram, graded by degree.

Lemma 2.6. *$\mathcal{P}_m$ is an associative algebra under concatenation, and the projection $\pi\colon \mathcal{D} \to \mathcal{P}_m$ is a homomorphism of algebras.*

Proof. Write $\mathcal{P}_m = \mathcal{D}/(R + I_m)$, where $R \subseteq \mathcal{D}$ is spanned by the relations of [3, Sect. 2.5] and $I_m \subseteq \mathcal{D}$ by the diagrams of degree $> m$. For the product to descend to the quotient it is enough that concatenating any diagram on either side of an element of $R + I_m$ again lies in $R + I_m$. For I_m this follows from the additivity of degrees under concatenation. For R it follows from Remark 2.4 as the concatenation attaches the new diagram at the basepoint and all relations remain relations. □

Finally, we recall the following theorem by Goussarov, Polyak and Viro which will be crucial for the attack.

Theorem 2.7 (Goussarov–Polyak–Viro [3])**.** *A knot invariant V with values in $\mathbb{Q}$ is a finite type invariant of degree m if and only if $V = f \circ \varphi_m$ for a linear functional $f\colon \mathcal{P}_m \to \mathbb{Q}$.*

3. The attack

In this section we recover the shared secret from the public and exchanged information of Alice and Bob. We begin by proving a few properties of the map φ_m.

Lemma 3.1 (Multiplicativity)**.** *For all knots K_1, K_2, we have*

$$\varphi_m(K_1 \# K_2) = \varphi_m(K_1)\,\varphi_m(K_2) \in \mathcal{P}_m.$$

Proof. Let D_1, D_2 be any diagrams representing the knots K_1, K_2 and form the standard connected sum diagram with basepoint on a connecting arc, so that its Gauss diagram is the concatenation $D_1 \cdot D_2$.

A subdiagram of $D_1 \cdot D_2$ is a choice of a subset of the arrows of $D_1 \cdot D_2$. Since no arrow of $D_1 \cdot D_2$ has one endpoint among those of D_1 and the other among those of D_2, such a subset decomposes uniquely as a subset S_1 of the arrows of D_1 together with a subset S_2 of the arrows of D_2. The based arrow diagram it spans is exactly the concatenation of the one spanned by S_1 with the one spanned by S_2. The signs multiply, so $\varepsilon(S_1 \cup S_2) = \varepsilon(S_1)\varepsilon(S_2)$. Therefore, we have in $\mathcal{D}$ that

$$\varphi(D_1 \cdot D_2) = \sum_{S_1, S_2} \varepsilon(S_1)\varepsilon(S_2)\,[S_1] \cdot [S_2] = \Big(\sum_{S_1} \varepsilon(S_1)[S_1]\Big) \cdot \Big(\sum_{S_2} \varepsilon(S_2)[S_2]\Big) = \varphi(D_1) \cdot \varphi(D_2).$$

Applying π, which is a homomorphism by Lemma 2.6, and using that φ_m is a knot invariant finishes the proof. □

Lemma 3.2 (Invertibility)**.** *For every knot K we have $\varphi_m(K) = \mathbf{1} + \nu$ with ν in the span of the diagrams of degree between 1 and m, and $\varphi_m(K)$ is invertible in $\mathcal{P}_m$ with*

$$\varphi_m(K)^{-1} = \mathbf{1} - \nu + \nu^2 - \cdots + (-1)^m \nu^m.$$

Proof. The empty subdiagram occurs in $\varphi(D)$ exactly once, so the degree-0 component of $\varphi_m(K)$ is $\mathbf{1}$. Write ν for the remainder, which lies in the span of diagrams of degree between 1 and m. Since degrees add and diagrams of degree $> m$ vanish in $\mathcal{P}_m$, we get $\nu^{m+1} = 0$ and therefore

$$(\mathbf{1} + \nu) \sum_{j=0}^{m} (-1)^j \nu^j = \mathbf{1} + (-1)^m \nu^{m+1} = \mathbf{1}.$$

□

Now we have established all of the crucial ingredients for the attack. While the monoid of oriented knots has no non-trivial invertible elements, we use that any finite type invariant factors through φ_m which maps the knots into the group of elements of $\mathcal{P}_m$ with degree-0 part $\mathbf{1}$.

Theorem 3.3. *Let V be any finite type invariant of degree m, and write $V = f \circ \varphi_m$ which is possible by Theorem 2.7. For all knots A, B, K, we have*

$$V(A \# B \# K) = f\big(\varphi_m(A \# K) \cdot \varphi_m(K)^{-1} \cdot \varphi_m(B \# K)\big).$$

Proof. By Lemma 3.1, $\varphi_m(A \# K) = \varphi_m(A)\varphi_m(K)$ and $\varphi_m(B \# K) = \varphi_m(B)\varphi_m(K)$. By Lemma 3.2 the factor in the middle is defined, so

$$\varphi_m(A \# K)\varphi_m(K)^{-1}\varphi_m(B \# K) = \varphi_m(A)\varphi_m(B)\varphi_m(K) = \varphi_m(A \# B \# K),$$

again by Lemma 3.1. After applying f the claim follows. □

Note that φ_m is evaluated on whichever diagrams are transmitted. Since it is a knot invariant the computation is independent of the "complexification" steps specified by Sconza and Wildi. The attack itself is now straightforward.

Algorithm 3.4 (Recovery of the shared secret)**.** Input: the public V (equivalently f), a diagram of K, and the transmitted diagrams $\widetilde{D}_A$ of $A \# K$ and $\widetilde{D}_B$ of $B \# K$.

(1) Compute the image under φ_m of the three input diagrams.
(2) Compute $\varphi_m(K)^{-1}$ using Lemma 3.2.
(3) Output $H\big(f(\varphi_m(\widetilde{D}_A)\varphi_m(K)^{-1}\varphi_m(\widetilde{D}_B))\big)$.

Lemma 3.5. *Algorithm 3.4 recovers the shared secret in time*

$$\widetilde{O}\big(c^{\lceil m/2 \rceil}\big) + \mathrm{poly}(\dim \mathcal{P}_m),$$

where c bounds the number of crossings of the three input diagrams to the algorithm. That is, an adversary can recover the shared secret at the cost of at most three honest evaluations of V, plus some linear algebra in dimension $\dim \mathcal{P}_m$.

Proof. The algorithm of [2], quoted as [7, Thm. 2.19], computes any finite type invariant of degree m in time $\widetilde{O}(c^{\lceil m/2 \rceil})$. In fact, it computes φ_m itself, each individual invariant being recovered from it by a linear functional, so obtaining the whole of φ_m costs no more than obtaining a single invariant of degree m. Step (1) applies this to three diagrams, and Steps (2) and (3) are arithmetic in the finite-dimensional algebra $\mathcal{P}_m$ and do not involve c. □

Note that increasing m or n raises the attacker's and the users' costs by the same factor.

4. The size of the key space

Aside from the attack presented in the previous section, the key space of Protocol 1.1 is too small for the parameters proposed in [7].

By Theorem 2.7 the invariant V is a fixed integer combination of finitely many of the counts $\langle \alpha, D \rangle$, and each such count ranges over subsets of at most m of the c crossings of the diagram, so $|\langle \alpha, D \rangle| \le \binom{c}{m}$, the degrees m of interest being far below $c/2$. Hence V takes $O(c^m)$ values on diagrams with c crossings, and the key $H(V(\cdot))$ carries at most

$$m \log_2 c + \log_2 C_V + 1 \quad \text{bits.}$$

Here, C_V is at most the number of based arrow diagrams of degree at most m times the largest coefficient of V. Suppose the coefficients of V are of size $O(1)$ and take $n \approx 10^3$, as proposed in [7, Sect. 4.4]. The diagrams the parties actually handle have c larger than n. Evaluated at

TABLE 1. Key space and cost of one evaluation of a degree-m invariant, at $c = 10^3$.

m	2	4	6	8	10
key space (bits)	25	52	81	110	140
cost of one evaluation	2^{10}	2^{20}	2^{30}	2^{40}	2^{50}

$c = 10^3$, the size of the key space and the cost of one evaluation of V by the algorithm of [2] are given in Table 1.

This means, to establish a key of 128 bits requires $m \geq 10$ and a single evaluation of a degree-10 invariant on a diagram of this size then costs of the order 2^{50} operations. Using larger knots does not counter this effect asymptotically either, as the space of possible shared keys grows only logarithmically in the crossing number whereas the cost grows polynomially. However, the space of secret knots grows exponentially in the crossing number [1].

The bound is moreover generous, in that it treats every integer in the range of V as attainable and equally likely. In experiments on knots small enough to enumerate exactly we found the realised key sets to be substantially smaller than it permits and not uniformly distributed.

5. Conclusion

Two parties holding different diagrams of the same knot can agree on a key only through a knot invariant, so the shared secret is necessarily $V(A \# B \# K)$ for some efficiently computable V. Theorem 3.3 breaks the proposed key exchange for all of them simultaneously. Invariants of another type were already surveyed in [7, Sect. 4.3] and ruled out, either because they admit a connected sum formula or because they cannot be computed efficiently.

Even ignoring the attack, little entropy survives the invariant. The key carries at most $m \log_2 c + \log_2 C_V + 1$ bits, with C_V as in Section 4, growing logarithmically in the crossing number, while evaluating it costs $\widetilde{O}(c^{\lceil m/2 \rceil})$, growing polynomially. Moreover, knots have no canonical representation, which leaves the concrete transmitted diagram as a further attack surface, independent of the one exploited here. For all these reasons, knots do not appear promising for instantiating a secure key exchange.

Acknowledgements

Claude Opus 5 was used to revise the text, improve flow, correct typos, generate the code used to draw Figure 1 and to carry out the computations mentioned in Section 4.

References

[1] I. Alekseev, A. M. Vershik and A. V. Malyutin. *On the growth of the number of prime knots.* Algebra i Analiz 36(1):17–39, 2024.

[2] D. Bar-Natan, I. Bar-Natan, I. Halacheva and N. Scherich. *Computing finite type invariants efficiently.* arXiv:2408.15942, 2024.

[3] M. Goussarov, M. Polyak and O. Viro. *Finite-type invariants of classical and virtual knots.* Topology 39(5):1045–1068, 2000.

[4] S.-P. Merz and C. Petit. *Factoring Products of Braids via Garside Normal Form.* Public-Key Cryptography – PKC 2019, Lecture Notes in Computer Science, 646–678, Springer, 2019.

[5] M. Polyak and O. Viro. *Gauss diagram formulas for Vassiliev invariants.* Internat. Math. Res. Notices 1994, no. 11, 445–453.

[6] J. Roberts. *Knots Knotes.* Lecture notes, Edinburgh course Maths 415, 1999; revised January 2015. https://mathweb.ucsd.edu/~justin/Roberts-Knotes-Jan2015.pdf

[7] S. Sconza and A. Wildi. *A Knot-based Key Exchange protocol.* Cryptology ePrint Archive, Paper 2024/471.