La véritable (et méconnue) théorie de l'information de Shannon

Olivier RIOUL[1], Julien BÉGUINOT[1], Victor RABIET[1,2] et Antoine SOULOUMIAC[3]

[1]LTCI, Télécom Paris, Institut Polytechnique de Paris, France [2]DMA, ENS Paris, France

[3]Université Paris-Saclay, CEA, LIST, F-91120, Palaiseau, France

nom.prenom@telecom-paris.fr, victor.rabiet@ens.fr, antoine.souloumiac@cea.fr

Résumé – Claude Shannon, qui a posé les bases de la célèbre théorie de l'information en 1948, préférait parler de théorie de la communication. Dans un article de 1953, il a proposé une nouvelle approche de l'information et de sa structure en treillis. Cela n'a étonnamment eu qu'un très faible écho. Nous rappelons brièvement les traits principaux de cette théorie et les caractéristiques qui pourraient en expliquer l'oubli, et nous envisageons enfin les développements intéressants qu'elle pourrait susciter.

Abstract – Claude Shannon, who laid the foundations of the celebrated information theory in 1948, preferred the term "communication theory". In a 1953 paper, he proposed a new approach to information and its lattice structure. Surprisingly, this remained relatively unknown. We briefly recall the features of this theory and the characteristics that could explain its poor impact; finally we consider possible interesting developments.

1 Introduction

« *Claude [Shannon] n'aimait pas le terme "théorie de l'information"* » nous rappelle Robert Fano, un collègue de Shannon travaillant au MIT, mort presque centenaire il y a à peine six ans. Dans une de ses dernières interviews [3], il précise : « *Vous comprenez, le terme "théorie de l'information" suggère que c'est une théorie sur l'information, mais ce n'est pas le cas. C'est sur la transmission d'information, pas l'information. Beaucoup de gens n'ont simplement pas compris cela.* »

Fano fait bien sûr référence à la célèbre théorie de Shannon de son article de 1948 [10] qu'il a intitulé « une théorie mathématique de la *communication* » – et non de l'information. Mais très tôt, c'est le terme « information » qui prévaut. L'entropie $H(X)$ d'une variable aléatoire discrète X est présentée comme la mesure « d'*information* contenue dans X », et la notion d'*information mutuelle* $I(X;Y)$ entre deux variables X et Y, introduite précisément par le même Robert Fano dans son cours au MIT [2], devient vite centrale dans l'enseignement de la théorie. D'ailleurs, le tout premier article historique sur la théorie, à peine 3 ans après sa naissance (!) s'intitule « *histoire de la théorie de l'information* » [1].

Cet engouement subit pour l'« information » aux début des années 1950 a fini par quelque peu ennuyer Shannon, qui, en 1956, dans son célèbre éditorial *The Bandwagon* [13] met en garde contre les dérives d'une telle popularité : « *Il sera trop facile pour notre prospérité un peu artificielle de s'effondrer du jour au lendemain lorsqu'on se rendra compte que l'utilisation de quelques mots excitants comme information, entropie, redondance, ne résout pas tous nos problèmes.* »

On peut, dans ces conditions, comprendre que Shannon ait voulu aller plus loin : si plusieurs variables aléatoires peuvent avoir la même *quantité* d'information H, comment définir l'information proprement dite ? Shannon présente le résumé de son analyse au congrès international des mathématiciens en 1950 [11] et dans un petit article relativement méconnu [12], publié en 1953 dans le tout premier numéro de ce qu'allait devenir les *IEEE Transactions on Information Theory*. Aujourd'hui, ce petit article n'a été cité que 153 fois [1] en comparaison des 139572 citations pour le grand article de 1948 [10].

La théorie de la « vraie information » (*actual information*) de Shannon est pourtant séduisante et il nous semble un peu injuste qu'elle soit resté confidentielle car elle est sans doute susceptible de développements intéressants. Cet article présente cette théorie en langage moderne, fournit des preuves (absentes chez Shannon), ainsi que quelques perspectives.

2 Qu'est-ce que l'information ?

2.1 Définition de la « vraie » information

L'idée de Shannon [12] est la suivante : si l'information contenue dans une source discrète X n'est pas la "mesure de quantité d'information" comme $H(X)$, ce doit être X elle-même ! Bien entendu, tout encodage réversible de X doit être considéré comme la *même* information, puisque que l'on passe d'une représentation à l'autre sans perte d'information. Cela revient, en langage moderne, à la définition suivante :

Définition 1 (« vraie » information)**.** *L'*information *(contenue dans) X est la classe d'équivalence de X selon la relation d'équivalence :*

$$X \equiv Y \iff Y = f(X) \text{ et } X = g(Y)$$

pour deux fonctions déterministes f et g.

Dire que $X \equiv Y$ revient encore à dire qu'il existe une bijection f telle que $X = f(Y)$. Dans la suite on notera encore sans

1. Et encore, il n'y a qu'une dizaine de citations correctes à dessein ; les autres sont en réalité des confusions avec l'article de 1948 [10] !

confusion possible X la classe d'équivalence de la variable X et donc $X = Y$ l'égalité entre les deux classes X et Y (plutôt que $X \equiv Y$).

Il est entendu que dans cette définition, on a un ensemble donné de variables aléatoires. Pour simplifier, considérons avec Shannon l'ensemble des variables discrètes X qui prennent un nombre *fini* de valeurs dans l'alphabet $\mathcal{X}$. Cela revient à considérer toutes les variables aléatoires $X : \Omega \to \mathcal{X}$ définies sur un espace probabilisé donné $(\Omega, \mathbb{P})$ où $|\Omega|$ est fini.

Avec cette définition, on voit clairement que la relation d'équivalence est compatible avec toute relation fonctionnelle $Y = f(X)$. Si f n'est pas bijective, on peut dire qu'il y a moins d'information en Y qu'en X. D'où l'ordre partiel suivant.

Définition 2 (Ordre partiel [2])**.**

$$X \geq Y \iff Y = f(X)$$

pour une fonction déterministe f.

On écrit aussi $Y \leq X$. C'est bien une relation d'ordre partiel car (1) $X \leq X$ (*réflexivité*, avec $f =$ identité) ; (2) $X \leq Y$ et $Y \leq X$ implique $X = Y$ (*antisymétrie*, par la définition 1) ; et (3) $X \leq Y$ et $Y \leq Z$ implique $X \leq Z$ (*transitivité*, par composition de fonctions).

2.2 Structure de treillis d'information

Au delà de l'ordre partiel, Shannon établit la structure mathématique naturelle de l'information : c'est un treillis (*lattice* en anglais) c'est à dire que deux variables X, Y admettent toujours un maximum $X \vee Y$ et un minimum $X \wedge Y$. Rappelons que ces quantités (nécessairement uniques si elle existent) sont définies par les relations [3]

$$(X \leq Z \quad \& \quad Y \leq Z) \iff X \vee Y \leq Z,$$
$$(X \geq Z \quad \& \quad Y \geq Z) \iff X \wedge Y \geq Z.$$

Proposition 1 (information totale)**.** *L'information totale* $X \vee Y$ *de* X *et* Y *est le couple aléatoire* $X \vee Y = (X, Y)$.

Démonstration. Que X et Y soient toutes deux fonctions de Z revient clairement à dire que (X, Y) est fonction de Z. □

La définition de l'information commune (*common information*) $X \wedge Y$ est plus ardue et non explicitée par Shannon. Suivant Gács et Körner [4] adoptons la définition suivante :

Définition 3. *On dit que* $x \in \mathcal{X}$ *et* $y \in \mathcal{Y}$ communiquent*, et on note* $x \sim y$*, s'il existe un chemin* $xy_1x_1y_2 \cdots y_nx_ny$ *dont toutes les transitions sont de probabilité non nulle :* $\mathbb{P}(X = x, Y = y_1) > 0, \mathbb{P}(Y = y_1, X = x_1) > 0, \ldots, \mathbb{P}(X = x_n, Y = y) > 0$.

C'est clairement une relation d'équivalence sur l'ensemble des couples (x, y) où $\mathbb{P}(X = x) > 0$ et $\mathbb{P}(Y = y) > 0$. Notons $C(x, y)$ la classe d'équivalence (*classe de communication*) de (x, y).

2. Nous ne considérons pas nécessairement des variables $X : \Omega \to \mathcal{X}$ réelles, et l'ordre $X \geq Y$ n'a rien à voir avec l'ordre dans $\mathbb{R}$.

3. Shannon utilise les notations booléennes $X + Y$ pour $X \vee Y$ et $X \cdot Y$ pour $X \wedge Y$.

Proposition 2 (information commune)**.** *L'information commune* $X \wedge Y$ *de* X *et* Y *est* $X \wedge Y = C(X, Y)$.

Démonstration. Que $Z = f(X) = g(Y)$ soit à la fois fonction de X et de Y revient à dire que Z est constante pour chaque couple (x, y) tel que $x \sim y$; autrement dit Z est une fonction de la classe $C(X, Y)$. □

On peut visualiser l'information commune de la manière suivante : la matrice stochastique définissant (X, Y) – la matrice des probabilités conjointes $\mathbb{P}(X = x, Y = y)$ – peut s'écrire après permutation éventuelle des lignes/colonnes sous la forme "bloc-diagonale" :

$$\mathbb{P}_{X,Y} = \begin{pmatrix} C_1 & & & \\ & C_2 & & \mathbf{0} \\ & & \ddots & \\ \mathbf{0} & & & C_k \end{pmatrix} \tag{1}$$

où le nombre de blocs k est maximal. Les k matrices rectangulaires représentent alors les k différentes classes d'équivalence, la probabilité $\mathbb{P}(C(X, Y) = i)$ étant la somme de tous les coefficients du bloc C_i.

Comme dans tout treillis, $X \leq Y$ revient à dire que $X \vee Y = Y$ ou que $X \wedge Y = X$.

2.3 Propriétés du treillis d'information

Proposition 3 (information nulle et information complète)**.** *Le treillis d'information est borné, i.e., il admet un élément minimum* 0 *et maximum* 1*, tel que pour tout* X*,* $0 \leq X \leq 1$.

Démonstration. L'information nulle 0 n'est autre que la classe de toute variable *déterministe*, qui vérifie clairement la relation $0 \leq X$ pour tout X. L'information complète 1 peut être vue comme le vecteur de toutes les variables considérées, ou ce qui revient au même pour des variables définies sur Ω, comme l'identité $1 : \Omega \to \Omega$. □

Proposition 4 (information manquante)**.** *Le treillis d'information est complémenté, i.e., tout* $X \leq Y$ *admet un* Z *tel que* $X \vee Z = Y$ *et* $X \wedge Z = 0$.

Ce Z est l'information manquante à X pour obtenir Y : elle permet de reconstruire Y à partir de X sans requérir plus d'information que nécessaire. Shannon ne dit pas comment la déterminer. Voici une construction possible.

Démonstration. Puisque $X \leq Y$, on a simplement $X = X \wedge Y = C(X, Y)$. Ainsi, une classe donnée $C(X, Y) = x$ n'y a qu'une seule valeur $X = x$ par classe, correspondant en général à plusieurs valeurs de Y, disons, $y_1^x, y_2^x, \ldots, y_{k_x}^x$. Posons alors $Z \in \{1, \ldots, k_X\}$ l'unique indice tel que $Y = Y_Z^X$.

Par construction $Z \leq X \vee Y = Y$, et puisque $X \leq Y$, il vient aussi $X \vee Z \leq Y$. Mais l'identité $Y = Y_Z^X$ montre que $Y \leq X \vee Z$, d'où l'égalité $X \vee Z = Y$.

Enfin la valeur $Z = 1$ connecte chaque paire (x, z), il n'y a donc qu'une seule classe selon (X, Z), c.-à-d. $X \wedge Z = 0$. □

On peut visualiser cette construction sur le tenseur stochastique de (X, Y, Z) décrit à la Fig. 1.

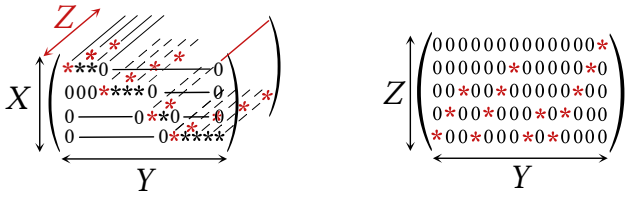

FIGURE 1 – Construction de l'information complémentaire Z permettant de passer de X à Y. On obtient la loi de Z en marginalisant le tenseur sur l'axe Z.

Notons, avec Shannon, que l'information manquante Z n'est *pas* uniquement déterminée par X et Y. Dans notre construction précédente, elle dépend de la façon d'indexer les valeurs de Y par classe $X = x$.

2.4 Une algèbre de Boole ?

C'est Shannon qui, dès 1938 dans son mémoire de master, avait montré l'utilisation de l'*algèbre de Boole* dans l'étude des circuits à base de relais – « le plus important mémoire de master du siècle » pour lequel Shannon a reçu le prix Alfred Noble [4] en 1940. Mais hélas, Shannon a dû ici constater que le treillis d'information n'est *pas* une algèbre de Boole. Ce serait le cas s'il était distributif ($\wedge$ distributif part rapport à $\vee$ ou l'inverse) car une algèbre de Boole est précisément un treillis borné complémenté distributif. Or, dans une algèbre de Boole, le complément est unique, ce qui n'est pas le cas ici (comme on l'a vu, l'information complémentaire n'est pas unique). Par conséquent, le treillis d'information n'est pas distributif.

3 Une métrique entropique. . .ou deux

3.1 Lien information-entropie

Il était évidemment important, pour Shannon, de vérifier la compatibilité du treillis d'information par rapport à des quantités comme l'entropie, qui mesure une quantité d'information. Tout d'abord, l'entropie (ou l'entropie conditionnelle, ou l'information mutuelle) est bien compatible avec la définition 1 de l'information comme classe d'équivalence, puisque deux variables en bijection ont essentiellement la même loi et donc la même entropie. On a ensuite des liens évidents :

Proposition 5 (ordre partiel et entropie conditionnelle)**.**

$$X \leq Y \iff H(X|Y) = 0$$

En particulier, H est "croissante" (une plus grande information a une plus grande entropie) :

$$X \leq Y \implies H(X) \leq H(Y)$$

et $H(X) \geq H(0) = 0$ pour tout X, avec égalité $H(X) = 0 \iff X = 0$.

4. À ne pas confondre avec le prix Alfred Nobel. . .

Démonstration. $H(X|Y) = 0$ signifie que $H(X|Y = y) = 0$ pour tout $y \in \mathcal{Y}$, ce qui revient à dire que X est déterministe $= f(y)$ sachant $Y = y$. Autrement dit $X = f(Y)$. On a alors bien $H(X) = H(X) - H(X|Y) = I(X;Y) = H(Y) - H(Y|X) \leq H(Y)$. Enfin l'entropie d'une variable est nulle si et seulement si cette variable est déterministe. □

Un exemple de treillis d'information avec les entropies associées est donné à la Fig. 2.

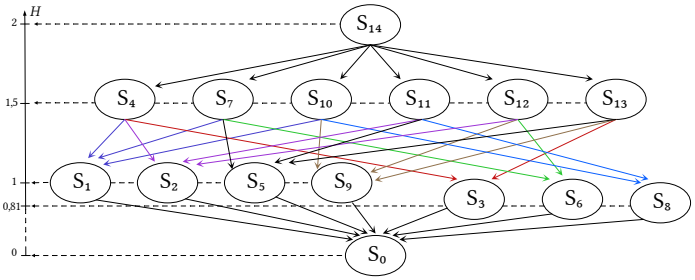

FIGURE 2 – Diagramme de Hasse du treillis d'information défini sur Ω de taille 4 muni de la probabilité uniforme. Les entropies correspondantes sont, par niveau descendant : 2 ; 1,5 ; 1 ; ≈0,81 et 0 bits.

3.2 Information commune et mutuelle

Il est évident que l'entropie de l'information conjointe est l'entropie conjointe $H(X \vee Y) = H(X, Y)$, et on peut se demander par analogie avec le diagramme de Venn usuel en théorie de l'information si l'entropie de l'information commune est égale à l'information mutuelle : est-il vrai que $H(X \wedge Y) = I(X;Y)$? La réponse est *non* :

Proposition 6. *$H(X \wedge Y) \leq I(X;Y)$ avec égalité si et seulement si on peut écrire $X = (U, W)$ et $Y = (V, W)$ où U et V sont indépendants sachant W.*

En particulier si X et Y sont indépendantes, on a bien une information commune nulle $X \wedge Y = 0$, mais en général $H(X \wedge Y) < I(X;Y)$.

Démonstration. Soit $W = X \wedge Y$. Puisque $W \leq X$ et $W \leq Y$, par complémentarité on peut écrire $X = W \vee U = (U, W)$ et $Y = W \vee V = (V, W)$. On a alors $I(X;Y) = I(U, W; V, W) = I(W; V, W) + I(U; V, W|W) = H(W) + I(U; V|W)$ ce qui démontre la proposition. □

Notons que le cas d'égalité correspond au cas où les blocs C_i de la matrice (1) sont des matrices stochastiques de deux variables indépendantes X, Y sachant $W = i$, c'est-à-dire des matrices de rang 1.

La proposition 6 est sous-jacente dans [4], et explicitée par Aaron Wyner dans [14] qui crédite une communication privée de Kaplan. À noter que l'information commune (*common information*) est plus connue aujourd'hui sous l'acception de Wyner, définie comme le maximum de $I(X, Y; W)$ lorsque X et Y sont conditionnellement indépendantes sachant W. Cette quantité est *supérieure* à l'information mutuelle et trouve son utilité en compression de sources corrélées.

3.3 Distances de Shannon et de Rajski

Puisque $X = Y \iff X \leq Y \ \& \ X \geq Y$, il suffit, d'après la Prop. 5, que $H(X|Y) + H(Y|X) = 0$ pour que $X = Y$. Shannon découvre là une *distance* qui rend le treillis d'information *métrique* [12] :

Proposition 7 (distance entropique de Shannon)**.** $D(X,Y) = H(X|Y) + H(Y|X)$ *est une distance.*

Démonstration. On a déjà noté que $D(X,Y) \geq 0$ ne s'annule que lorsque $X = Y$. La symétrie $D(X,Y) = D(Y,X)$ est évidente. L'inégalité triangulaire $D(X,Z) \leq D(X,Y) + D(Y,Z)$ provient de l'inégalité $H(X|Z) \leq H(X,Y|Z) = H(X|Y,Z) + H(Y|Z) \leq H(X|Y) + H(Y|Z)$, sommée avec l'inégalité obtenue en permutant X et Z. □

Il est intéressant de noter que ce n'est pas la seule distance (ni la seule topologie). En normalisant $D(X,Y)$ par l'entropie conjointe $H(X;Y)$, on obtient encore une distance :

Proposition 8 (distance de Rajski [8])**.** $d(X,Y) = \frac{D(X,Y)}{H(X,Y)}$ *est une distance*[5] *(toujours comprise entre* 0 *et* 1*).*

Noter que $\rho(X,Y) = 1 - d(X,Y) = \frac{I(X;Y)}{H(X,Y)}$ est la définition naturelle d'un coefficient de dépendance[6] (nulle si X et Y sont indépendantes, égale à 1 si X et Y sont équivalentes).

Démonstration. Il suffit de vérifier l'inégalité triangulaire. La preuve est identique à la vérification similaire sur l'index de Jaccard/Tanimoto vis-à-vis de la différence symétrique. Une preuve simple se trouve dans [5]. □

3.4 Une théorie discontinue

Le plus grand défaut de la « vraie » théorie de l'information de Shannon semble être que les différentes constructions d'éléments dans le treillis (information commune, complémentaire) ne dépendent pas, en réalité, des *valeurs* des probabilités en jeu mais seulement du fait qu'elles sont égales ou différentes de zéro. Ainsi, une petite perturbation sur des probabilités peut grandement influer sur les résultats. Par exemple :

Proposition 9 (Discontinuité de l'information commune)**.** *L'application* $(X,Y) \mapsto X \wedge Y$ *est* discontinue *dans le treillis métrique muni de la distance* D *(ou* d*).*

Démonstration. $(X_\varepsilon, Y_\varepsilon)$ définie par la matrice stochastique

$$\mathbb{P}_{X,Y} = \begin{pmatrix} \frac{1-\varepsilon}{N} & \frac{\varepsilon}{N} & 0 & \cdots & 0 \\ 0 & \frac{1-\varepsilon}{N} & \frac{\varepsilon}{N} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots \\ \frac{\varepsilon}{N} & 0 & \cdots & 0 & \frac{1-\varepsilon}{N} \end{pmatrix}.$$

a pour information commune $X_\varepsilon \wedge Y_\varepsilon = 0$ si $\varepsilon > 0$ (une seule classe de communication) alors que $X_\varepsilon \wedge Y_\varepsilon$ est uniformément distribuée sur N classes si $\varepsilon = 0$. Par conséquent $D(X_\varepsilon \wedge Y_\varepsilon, 0) = 0$ pour tout $\varepsilon > 0$ alors que $D(X_0 \wedge Y_0, 0) = H(X_0 \wedge Y_0) = \log N$ est arbitrairement grand pour $\varepsilon = 0$. □

5. Par convention $d(0,0) = 0$.
6. On pourrait faire un parallèle avec le coefficient de corrélation.

4 Conclusion et perspectives

Il est peu de dire que la « véritable » théorie de l'information de 1953 n'a pas eu le même succès que celle de 1948. John Pierce, également collègue de Shannon, le commente ainsi [7] : « *Apparemment, la structure n'était pas assez bonne pour mener à quelque chose de grande valeur.* » Nous y trouvons deux raisons possibles : le fait que le treillis ne soit pas booléen, ce qui ne facilite pas les calculs ; et surtout, le caractère discontinu du treillis vis-à-vis de la métrique entropique.

Certaines applications sont néanmoins intéressantes. Récemment, [6] a utilisé le treillis d'information pour évaluer les fuites d'information dans l'exécution de programmes déterministes.

Dans une optique plus proche du traitement du signal, en explicitant les conditions pour que trois variables X, Y, Z soient *alignées* (vérifient l'inégalité triangulaire avec égalité), on peut résoudre (ou non) des problèmes de reconstruction parfaite en fonction de relation sur les coefficients de dépendance [9, Exercice 6 p. 242]. Ce type de problème gagnerait à être généralisé.

Références

[1] E. C. Cherry, "A history of the theory of information," *Proc. Inst. Electrical Engineering*, vol. 98, no. 383–393, 1951.

[2] R. M. Fano, *Class notes for course 6.574 : Transmission of Information*, MIT, Cambridge, MA, 1952.

[3] ——, "Interview by Aftab, Cheung, Kim, Thakkar, Yeddanapudi, 6.933 Project History, Massachusetts Institute of Technology," Nov. 2001.

[4] P. Gács and J. Körner, "Common information is far less than mutual information," *Problems of Control and Information Theory*, vol. 2, no. 2, pp. 149–162, Jan. 1973.

[5] Y. Horibe, "A note on entropy metrics," *Information and Control*, vol. 22, no. 4, pp. 403–404, May 1973.

[6] P. Malacaria, "Algebraic foundations for quantative information flow," *Mathematical Structures in Computer Science*, vol. 25, no. 2, pp. 404–428, Feb. 2015.

[7] J. R. Pierce, "The early days of information theory," *IEEE Transactions on Information Theory*, vol. 19, no. 1, pp. 3–8, Jan. 1973.

[8] C. Rajski, "A metric space of discrete probability distributions," *Information and Control*, vol. 4, no. 4, pp. 371–377, Dec. 1961.

[9] O. Rioul, *Théorie de l'information et du codage*. Hermes Science - Lavoisier, 2007.

[10] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3 & 4, pp. 379–423 & 623–656, July & Oct. 1948.

[11] ——, "Some topics on information theory," in *Proc. Int. Congress Math.*, AMS, Ed., vol. II, Aug. 30 - Sept. 6 1950, pp. 262–263.

[12] ——, "The lattice theory of information, in Report of Proc. Symp. Inf. Theory, London, Sept. 1950," *Trans. IRE Professional Group Inf. Theory*, vol. 1, no. 1, pp. 105–107, Feb. 1953.

[13] ——, "The bandwagon (editorial)," *IRE Transactions on Information Theory*, vol. 2, no. 1, pp. 3–3, Mar. 1956.

[14] A. D. Wyner, "The common information of two dependent random variables," *IEEE Transactions on Information Theory*, vol. 21, no. 2, pp. 163–179, Mar. 1975.