Do Rational Equivalence Relations have Regular Cross-Sections? †

J. H. Johnson

Department of Computer Science
University of Waterloo
Waterloo, Ontario, Canada N2L 3G1

ABSTRACT

The following classes of rational equivalence relations are shown to have regular cross-sections: deterministic rational equivalence relations, rational equivalence relations over a one letter alphabet, and rational equivalence relations with bounded separability. Although the general case remains open, it is shown to be reducible to that of locally-finite rational equivalence relations over a two letter alphabet. Two particular cross-sections are shown not to be regular: the set of minimum length words and the set of lexicographically minimal words.

1. Introduction

The familiar Soundex code maps surnames into four-character codes so that similar sounding names are mapped to the same code and different sounding names are mapped to different codes. Thus "Johnson" and "Jansen" are both transformed into the code "J525" indicating an initial letter "J", a letter in the class $\{m,n\}$, a letter in the class $\{c,g,j,k,q,s,x,z\}$, and a letter in the class $\{m,n\}$. It is well known that Soundex is not perfectly accurate in its coding. For example, the common variant spelling "Johnston" is assigned a code "J523". The class "L000" contains all names with a initial "L" and any sequence of vowels and the letters "h" and "w". Many easily distinguishable Chinese surnames fall in this class. For a detailed description of Soundex and a discussion of its use and problems see [10, 12].

In order to improve on Soundex, a number of attempts have been made to construct coding functions that more closely reflect the structure of surnames in the population of interest. For example, the NYSIIS code was designed to achieve greater accuracy on a population with a significant fraction of Spanish surnames. A number of coding functions, including NYSIIS, are discussed by Moore *et al* [11].

One interesting observation [9] is that both the Soundex and NYSIIS functions can be usefully modelled as deterministic GSM's (subsequential functions). This means that they can be computed deterministically from left to right in one pass. It follows immediately that both Soundex and NYSIIS are rational functions and that the relation "has the same code" is a rational equivalence relation on surnames for either Soundex or NYSIIS. Rational relations are discussed by Berstel [1] and Eilenberg [4].

† This work was supported by the Natural Sciences and Engineering Research Council of Canada, Grant No. A0237

The advantage of these coding functions comes primarily from their low cost. Even very large files can be partitioned cheaply according to a code value using any of a number of sorting or hashing algorithms. The coding function needs to be computed only once and so a moderate amount of effort can be directed into computing better codes if the accuracy can be improved. The use of these types of coding functions has been advocated by various authors [3,6,10,11,12,17]. Thus more general equivalence relation models for which reasonable canonical functions are available are worthy of consideration.

It can be shown [9] that any rational equivalence relation has a canonical function computable in $O(n^2)$ time and space or in $O(n^3)$ time and $O(n)$ space where n is the length of the input. If the rational equivalence relation has a rational canonical function, however, it can be computed in $O(n)$ time and space resulting in a significant saving [13,15]. An interesting question is then to identify when a rational equivalence relation has a rational canonical function. It appears that all of them do but seems to be very difficult to prove. In fact, the question is still open. This paper will discuss the current state of this conjecture which is equivalent to that stated in the title.

2. Terminology

Definition: A (binary) **relation** over sets S and T is a subset of $S \times T$.

We will be interested in the usual Boolean operations on relations interpreted as sets as well as the operations composition, inversion (interchange of components), domain, range, identity, cross-product, and application:

$$R_1 \circ R_2 = \{(u,w) \mid \exists v[(u,v)\in R_1 \text{ and } (v,w)\in R_2]\} \qquad R^{(-1)} = \{(v,u) \mid (u,v)\in R\}$$

$$dom R = \{u \mid (u,v)\in R\} \qquad ran R = \{v \mid (u,v)\in R\} \qquad \iota_L = \{(u,u) \mid u\in L\}$$

$$L_1 \times L_2 = \{(u,v) \mid u\in L_1, v\in L_2\} \qquad R(L) = \{v \mid \exists u[u\in L \text{ and } (u,v)\in R]\}.$$

Definition: An **equivalence relation** R over a set S is a relation satisfying the reflexive, symmetric, and transitive laws: $\iota_S \subseteq R$, $R^{(-1)} \subseteq R$, and $R \circ R \subseteq R$.

Definition: The **kernel** of a function $f: S \to T$ is the equivalence relation over S:

$$ker f = \{(u,v)\in S \times S \mid f(u) = f(v)\} = f \circ f^{(-1)}.$$

Definition: A **canonical function** for an equivalence relation R over S is any function $f: S \to T$ satisfying $R = ker f$.

Definition: A **cross-section** of an equivalence relation R over S is a set D containing one element from each class of R. Then $f = R \cap (S \times D)$ is a canonical function.

Definition: The **restriction** $R \mid D$ of an equivalence relation R to a set D is the relation formed from R by restricting the domain and range to D. Thus $R \mid D = R \cap (D \times D)$.

Definition: A **thinning** of an equivalence relation R is a restriction whose domain contains at least one member of each equivalence class of R.

Definition: A relation R is **locally-finite** if for any $x \in dom R$, the set $\{y \mid (x,y)\in R\}$ is finite. If R is an equivalence relation then local-finiteness requires that every class be finite in size.

Definition: A **monoid** $<M, \cdot, \square>$ is a set M with an associative binary operation $\cdot$ and an identity element $\square$ satisfying: $a\cdot(b\cdot c) = (a\cdot b)\cdot c$, $\square \cdot a = a = a \cdot \square \quad \forall a,b,c \in M$.