

# How Consistent are your Choice Data?\*

Mark Dean and Daniel Martin<sup>†</sup>

Center for Experimental Social Science - New York University

*Preliminary and incomplete.*

November 8, 2008

## Abstract

This paper describes a fast and exact method for computing the Houtman-Maks (HM) Index - the largest subset of a choice data set consistent with acyclicity. This measure provides a metric for determining how close a set of choice data are to ‘rationality’, in the sense that they can be modeled as having been derived from the maximization of a complete preference relation. One reason that this measure has not been widely used is that it is extremely computationally intensive. We show that the problem of finding the maximal acyclic subset is isomorphic to a well-studied problem within computer science: the Minimum Set Covering Problem (MSCP). While the MSCP is NP-hard in the strong sense, there are a wide variety of algorithms built to solve this problem quickly and exactly for reasonably-sized data sets. This paper describes some of these algorithms and presents simulation results to demonstrate that our algorithm can be used to calculate the HM Index in under a second for cases that have previously been found insoluble.

---

\*We are grateful to Shachar Kariv for providing the data and computer program used in Choi et al [2006], Hiroki Nishimura for helpful comments, Jesse Perla for programming suggestions and Adam Sachs for research assistance.

<sup>†</sup>Department of Economics, New York University, 19 W. 4th St., 6th Floor, New York, NY 10012 (correspondence email: daniel.martin@nyu.edu).

# Introduction

Economists have a long tradition of specifying their models in terms of *axioms*, or restrictions on data which must be obeyed for it to be consistent with a particular model. One common axiom is that a binary relation (say  $\succ$ ) on some set  $Z$  must be *acyclic* - meaning that one cannot find a sequence  $z_1, z_2, \dots, z_n$  in  $Z$  such that

$$z_1 \succ z_2 \succ \dots \succ z_n \succ z_1$$

Acyclic, antisymmetric relations have the important property that they can be extended to complete linear orders.<sup>1</sup> As such, we can interpret  $\succ$  as being an incomplete observation of some consistent ordering on  $Z$ . The most famous use of this result is Strong Axiom of Revealed Preference (SARP), which shows that a set of choice data can be considered consistent with the maximization of some underlying preference ordering if and only if the generated revealed preference relation is acyclic. In this case, the relationship  $x \succ y$  is used to represent the observation that  $x$  is chosen when  $y$  is available. Caplin and Dean [2008] provide an example in which the acyclic property is central to characterising a model of search and choice.

One problem with the axiomatic method of characterizing a model is that it provides only a very stark measure of whether a data set is consistent with a particular model: data either does or does not violate the stated axioms. There is no concept of whether a data set is ‘close’ to satisfying an axiom set. Recognising this problem, several authors have proposed measures of how ‘far away’ a data set is from satisfying a set of axioms (see Afriat [1972], Varian [1991], and Houtman and Maks [1985]).

The Houtman and Maks measure is based on finding the largest subset of observations which satisfy the axiomatic system. For example, consider a choice experiment in which a subject exhibits the following behavior:

$$C(\{x, y\}) = \{x\}$$

$$C(\{x, y, z\}) = \{z\}$$

$$C(\{x, z\}) = \{z\}$$

$$C(\{y, z\}) = \{y\}$$

$$C(\{x, y, w\}) = \{w\}$$

---

<sup>1</sup>i.e., there exists a complete, antisymmetric, reflexive and transitive order  $\succ^*$  on  $Z$  such that  $x \succ y \Rightarrow x \succ^* y$

These data are not consistent with SARP, as  $x$  is revealed preferred to  $y$ , while  $y$  is revealed preferred to  $z$ , which is in turn revealed preferred to  $x$ . However, if one were to remove the observation  $C(\{y, z\}) = \{y\}$ , then the resulting system would be consistent with SARP. Alternatively, consider the following choice data:

$$\begin{aligned}
 C(\{x, y\}) &= \{x\} \\
 C(\{x, y, z\}) &= \{z\} \\
 C(\{x, z\}) &= \{z\} \\
 C(\{y, z\}) &= \{y\} \\
 C(\{x, y, w\}) &= \{y\}
 \end{aligned}$$

This data set is also not consistent with SARP. However, one would have to remove *two* observations before finding a set which is consistent. One might therefore say that the second set of choices is further away from consistency than is the first set. While this measure is not without its problems (see Choi et al. [2006] for a discussion), it has the advantage of being applicable to wide variety of data sets and axiomatic systems. In contrast, the Afriat and Varian measures are only applicable to data obtained by observing choices derived from different budget sets.

One possible reason that the Houtman and Maks measure has not been widely adopted is that it is extremely computationally intensive (see Choi et al. [2007] and Fisman et al. [2007] for examples in which computational constraints have been binding). While some papers (e.g., Gross and Kaiser [1996]) describe computational algorithms which solve similar problems, these methods tend to be either slow and exact or fast and inexact. The innovation in this paper is to show that the problem of finding the maximal acyclic subset is isomorphic to an extremely well-studied problem within the computer sciences and operations research: the *Minimum Set Covering Problem* (MSCP). While MSCP is NP-hard in the strong sense, there are a wide variety of algorithms built to solve this problem (see Caprara, Toth, and Fischetti [2000]), which can be used to find the maximal acyclic set quickly and exactly for reasonably-sized data sets. This paper describes some of these algorithms, and a companion website ([www.danielmartin.com/HM](http://www.danielmartin.com/HM)) provides code which adapts them to the Houtman-Maks measure. We demonstrate that it can be calculated in under a second for cases that have previously been found insoluble. This result opens up the possibility that the Houtman-Maks measure can be developed into a more

formal statistical test of axiomatic consistency.

## Basics

In this section we describe notation, set up the problem and show its equivalence to MSCP. Throughout, we illustrate the problem using the case of SARP from consumer theory.

Let  $Z$  be some arbitrary finite set (in the consumer theory example, let  $Z$  be the underlying options which the decision maker may choose among) and  $X$  be some (finite) set of observations (e.g., the result of various choice experiments - that  $z_1$  was chosen from the set  $Z_1 \subset Z$ ,  $z_2$  from  $Z_2 \subset Z$  and so on, represented as a tuple  $\{z_i, Z_i\}$ ). Let  $D : X \rightarrow 2^{Z \times Z}$  be the binary relations on  $Z$  generated by each observation in  $X$ . In our previous example,  $x_i = \{z_i, Z_i\}$  would generate the binary relation that  $z_i$  is preferred to all the other objects in  $Z_i$ . Thus  $D(\{z_i, Z_i\}) = \{(z_i, z) \in Z \times Z \mid z \in Z_i \setminus \{z_i\}\}$ .<sup>2</sup>

Denote by  $\succ_x$  the binary relation generated by observation  $x \in X$ , so that  $\succ_x = D(x)$ . For any  $A \subset X$ , define the binary relation  $\succ^A$  on  $Z$  as

$$z \succ^A w \text{ if, for some } x \in A, z \succ_x w$$

Thus, a set of observations  $X$  is consistent with SARP if the relation  $\succ^X$  is acyclic. More generally, we define a set  $A \subset X$  as acyclic if the binary relation it generates  $\succ^A$  is acyclic.

We are now in a position to define the Houtman-Maks measure for how close  $X$  is to defining an acyclic relation, which we will call the *HM Index*.

**Definition 1** *The HM Index for a set  $X$  is a number  $M$  such that  $M = |A|$ , where  $A$  is a maximal acyclical*

---

<sup>2</sup>Note that we require the binary relations generated by an observation  $x \in X$  be independent of whether some other observation  $y \in X$  is made.

set, defined as

$A \subset X$  such that

(i)  $A$  is acyclic

(ii)  $\forall B \subset X$  s.t.  $|B| > |A|$ ,

$\exists z_1, z_2, \dots, z_n \in Z$  s.t.

$z_1 \succ^B z_2 \succ^B \dots \succ^B z_n \succ^B z_1$

Next, we define a minimum covering problem. To give an idea of this class of problems, consider the following example: Say you are setting up a cell-phone network and need to buy rights to bandwidth in all 50 states. However, bandwidth is being sold in packages of different states (e.g., package 1 includes Alabama, Rhode Island and Wyoming, package 2 includes South Dakota, Minnesota and Wyoming and so on). Each package has a particular cost. The problem you face as a cell-phone provider is: ‘What collection of packages should I buy to ensure some bandwidth in all 50 states at the lowest possible cost?’ In other words, what is the minimum cost way of covering all 50 states? A formal statement of this class of problems is as follows <sup>3</sup>.

**Definition 2** Let  $S$  be some base set,  $\Theta$  be a collection of subsets  $\theta$  of  $S$ , and  $k : S \rightarrow \mathbb{R}$  be a cost function which attaches a cost to each element of  $S$ . A covering of  $\Theta$  is a subset  $T \subset S$  such that  $\theta \cap T \neq \emptyset \forall \theta \in \Theta$ . In other words, every set in  $\Theta$  contains at least one element of  $T$ . The minimum covering of  $\Theta$  is the solution to the problem

$$\min_{T \subset S} \sum_{x \in T} k(x)$$

subject to  $\theta \cap T \neq \emptyset \forall \theta \in \Theta$

In the bandwidth example above, we can let  $S$  be the set of packages and  $\Theta$  be a collection of 50 sets, one for each state, each containing the packages which cover that state (e.g., if Alabama was covered by packages 1, 7 and 9 then  $\theta_1 = \{1, 7, 9\}$ , if Alaska was covered by packages 3, 14, 19 and 23 then  $\theta_2 = \{3, 14, 19, 23\}$  and so on).  $k$  would contain information on the cost of each package.

---

<sup>3</sup>Note that some people call the problem stated this way as the ‘Minimum Hitting Problem’. However, Ausiello et al. [1980] show that this is equivalent to other statements of the Minimum Set Covering Problem.

To show that finding the HM Index can be reduced to a minimum covering problem, we introduce the concept of a *cycle*  $c$  generated by  $X$ .

**Definition 3** A *cycle* consists of a non-repeating sequence  $z_1, \dots, z_n$  in  $Z$  and a sequence  $x_1, \dots, x_n$  in  $X$  such that

$$z_1 \succ_{x_1} \dots \succ_{x_{n-1}} z_n \succ_{x_n} z_1$$

Let  $C$  denote the set of all cycles generated by  $X$ . We will say that removing an element  $x \in X$  *breaks* a cycle  $c$  if  $x$  appears in the sequence  $x_1, \dots, x_n$ .

We claim that the problem of finding the HM Index of  $X$  is the same as finding the minimal set  $B \subset X$  such that, between them, the elements of  $B$  break all cycles in  $C$ . In particular,  $M = |X| - |B|$ .

**Theorem 1** For a set  $X$ , the HM Index  $M$  is equal to  $|X| - |B|$ , where  $B$  is the smallest subset of  $X$  which breaks all cycles in  $C$ .

**Proof.** We rule out two cases by contradiction:

1.  $M > |X| - |B|$ . Let  $A$  be a maximal acyclical subset of  $X$  and consider the set  $B^* = X \setminus A$ . It must be true that  $B^*$  breaks all cycles in  $C$ . If not, then there exists some cycle  $\{z_1, \dots, z_n, x_1, \dots, x_n\}$  such that  $x_i \in A \forall i \in \{1, \dots, n\}$ . This in turn implies that

$$\begin{aligned} z_1 \succ_{x_1} \dots \succ_{x_{n-1}} z_n \succ_{x_n} z_1 \\ \Rightarrow z_1 \succ^A \dots \succ^A z_n \succ^A z_1 \end{aligned}$$

and so  $A$  would not be acyclic. But, as  $M = |A|$ , this implies that  $|A| > |X| - |B|$ , and so  $|X| - |B^*| > |X| - |B|$  and thus  $|B^*| < |B|$ , contradicting the fact that  $B$  is the smallest subset to break all cycles in  $C$ .

2.  $M < |X| - |B|$ . Let  $A^* = X \setminus B$ . We claim that  $A^*$  must be acyclic. If not, there exists some cycle  $z_1 \succ^A \dots \succ^A z_n \succ^A z_1$ . But this implies there exists some cycle  $\{z_1, \dots, z_n, x_1, \dots, x_n\}$  with all elements  $x_i \in A$ . This cycle is not broken by  $B$ , contradicting the claim that  $B$  breaks all cycles in  $C$ . But, as

$|A^*| = |X| - |B| > M$ , this contradicts the claim that  $M$  is equal to the size of the maximal acyclical subset.

■

Using the above lemma, it is clear that the problem of finding  $M$  can be couched as a minimum covering problem. Let  $S = X$ , the set of observations, and construct  $\Theta$  using the mapping

$$F : C \Rightarrow X$$

$$\text{such that } F(c) = \{x \in X \mid x \text{ breaks } c\}$$

and let  $\Theta = F(C)$ . Finally, by setting  $k(x) = 1 \forall x \in X$ , the problem of finding the smallest subset of  $X$  which breaks all cycles in  $C$  can be described as finding the minimum covering of  $\Theta$ .

## Method

Calculating the HM Index as a minimum covering problem requires two algorithmic components. First, to fully specify the mapping  $F$  on  $C$ , we need an algorithm that identifies the set of all cycles  $C$  and the choices  $x \in X$  that break each cycle  $c \in C$ . One option is Johnson’s Algorithm, a computationally efficient graph theory algorithm (see Johnson [1975]). A graph  $G$  is composed of nodes  $N$  and edges  $E$ , and preferences can be represented as a directed graph by creating a node for object ( $E = Z$ ) and placing a direction edge between nodes when one object is preferred to another (e.g.,  $e_1 = (z_1, z_2)$  if  $z_1 \succ z_2$ ). Johnson’s Algorithm is based on ‘depth-first’ search, a standard approach to finding cycles, which looks at the objects preferred to an initial object, then looks for the objects that are preferred to the first of those preferred objects and so on until a cycle is found or the process terminates. At that point, the algorithm goes back one level and proceeds from the second preferred object until all possibilities are exhausted.

To gain efficiency, Johnson adds a blocking function to prevent redundant searching on the tree, which gives it a computation time upper bound of  $O((n + e)(c + 1))$ , where  $n$  is the number of nodes,  $e$  is the number of edges and  $c$  is the number of cycles. We add additional efficiency by modifying Johnson’s Algorithm to only look

at those cycles without subcycles. For example, if  $a \succ_x b \succ_y a$  and  $a \succ_x b \succ_y c \succ_z a$ , then removing  $x$  or  $y$  would break both cycles, so breaking the first cycle is equivalent to breaking both cycles.

Second, we need an algorithm to solve the Minimum Set Covering Problem (MCSP), which is NP-hard. A problem is *NP* if the proof of a 'yes' answer can be verified in polynomial time, and a problem is *NP-hard* if every problem in NP can be reduced to that problem. Further, if a problem is in the set of NP problems and every problem in NP can be reduced to it, then that problem is *NP-complete*.

Determining whether a problem is NP-hard or NP-complete is typically accomplished by showing that is reducible to a single established NP-complete problem, and a problem  $G$  is reducible to a problem  $H$  if any version  $g \in G$  of that problem can be transformed into a version  $h \in H$  of the other problem. Our definition of MSCP includes an optimization problem that is not in the set of NP problems because the answer is not yes or no, but the problem can be reduced to a corresponding NP-complete problem which asks if there exists a set  $T$  of size  $n$  or less that covers  $\Theta$  (see Garey and Johnson [1979]). As a result, there does not currently exist an algorithm that solves either problem in polynomial time for all data sets.<sup>4</sup>

However, MSCP has been studied exhaustively because it can be applied to many real-world situations, such as train scheduling and city planning. As a result, algorithms have been developed to solve or approximate solutions to MSCP quickly for larger and larger data sets. Branch and bound algorithms find an exact solution by iteratively 'relaxing' the integer programming problem so that linear programming techniques can be used to create bounds on the problem. These algorithms have been integrated into standard Integer Programming (IP) software packages, including commercial solvers (e.g., CPLEX) and non-commercial solvers (e.g., SCIP and MINTO), which outperform stand-alone algorithms and tend to work quickly for most data sets (see Caprara, Toth, and Fischetti [2000]). In the following benchmark comparison, we use the `bintprog` command in the MATLAB Optimization Toolbox to solve MSCP.

---

<sup>4</sup>If you can write an algorithm that solves this NP-complete problem in polynomial time for all data sets, then you can claim a \$1 million prize from the Clay Institute!

## Benchmark

To benchmark our approach, we use the data and computer program from Choi et al [2006]. In this paper, 93 subjects allocate tokens between account x and account y using a novel graphical interface. Over 50 rounds, subjects face randomly selected budget lines, with a constant, normalized wealth level. The state of the world is uncertain, and in one state of the world, each x token pays \$.50 and each y token pays nothing, and in other state of the world, the reverse is true. Clearly, a bundle  $(x_1, y_1)$  is revealed preferred to a bundle  $(x_2, y_2)$  if both are available in the budget set and  $(x_1, y_1)$  is selected instead of  $(x_2, y_2)$ .

Choi et al [2006] report that calculating the HM Index with their approach is infeasible for subjects with data that is far from acyclicity. To see why this calculation becomes increasingly difficult, imagine a data set that has 50 observations and that removing 9 observations is necessary to make the data set consistent. Before you can determine that removing 9 observations is necessary, you must first check that removing any single observation is not sufficient, then any two observations and so on until you have checked all possible subsets of observations of size 8, which involves the following number of checks.

$$\binom{50}{49} + \binom{50}{48} + \binom{50}{47} + \binom{50}{46} + \binom{50}{45} + \binom{50}{44} + \binom{50}{43} + \binom{50}{42} > 655,000,000$$

To calculate the HM Index more quickly, the authors first partition their choice data using the strongly connected components of the corresponding graph, and then look at the minimum number of removals needed to make each partition consistent. This approach is similar to the exact approach detailed in Gross and Kaiser [1999] for undirected graphs, which focuses on disconnected subgraphs. In addition to reducing computation time, partitioning sets allows Choi et al [2006] to calculate an upper-bound on the minimum number of removals needed to make a subject's data consistent.

Even though most subjects were close to being perfect rational, Choi et al [2006] found that they were not able to produce an exact HM Index score for 5% (6 of 93) of subjects, who are listed in Table 1. Our approach, based on MSCP, was able to solve even the most difficult case in under a quarter of a second. Additionally, the benchmark lower bound was often significantly smaller than the size of the largest acyclical set. Both programs were run on a desktop computer with a 2 GHz dual core processor and 1.95 GB of RAM running Windows XP

Professional. The program for the our approach is available at [www.danielmartin.com/HM](http://www.danielmartin.com/HM).

Table 1

Benchmarking Results				
ID	Benchmark Approach		New Approach	
	HM Index (Lower Bound)	Run Time	HM Index (Exact)	Run Time
211	34	> 1 hr	35	0.1156 secs
324	29	> 1 hr	43	0.0475 secs
325	32	> 1 hr	41	0.0709 secs
406	30	> 1 hr	40	0.0451 secs
504	33	> 1 hr	45	0.0401 secs
608	29	> 1 hr	40	0.0451 secs

The computational difficulty of calculating the HM Index also meant that Choi et al. [2006] were not able to benchmark their results against random choice. Bronars [1987] discusses the role that random choice can play in determining the power of an index. To apply such an idea to the HM Index, one must be able to calculate it for random choices, which can be very far from rationality. The improved efficiency of our approach allows for such benchmarking. Figure 1 shows the distribution of HM Index scores for the subjects' choice data (reproduced from Choi et al. [2006]) and for hypothetical subjects selecting baskets of goods at random from the budget line. This simulation was run for 25,000 subjects that make uniformly distributed choices on randomly generated budget lines in keeping with the experimental design.

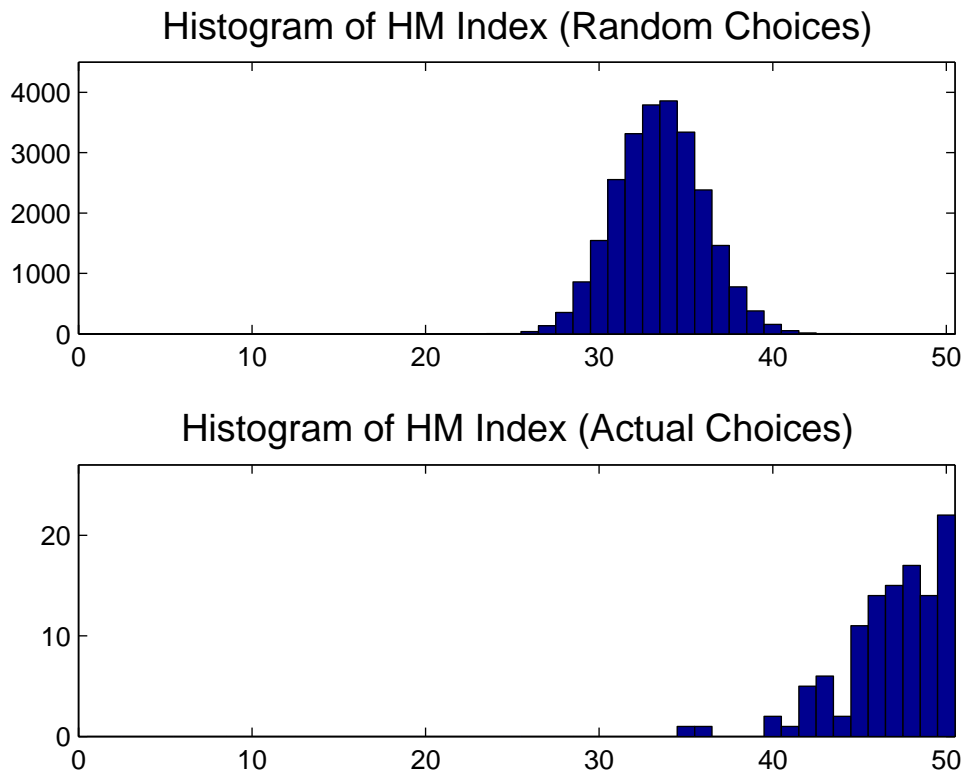


Figure 1: Comparison with Random Choice Data

## Conclusion

This paper provides a new tool to help answer the question ‘How close is a set of choice data to rationality?’. While not perfect, the HM Index is a flexible and powerful tool for answering this question, but has been largely abandoned on the basis that its computational difficulty. By showing the equivalence of calculating the HM index to MSCP, we have essentially removed these constraints. We can solve problems in fractions of a second that were previously insoluble in any reasonable length of time. By doing so, we have opened the door to more sophisticated use of the HM Index, using simulation and benchmarking against other models of choice.

## References

Afriat, S. (1972) "Efficiency Estimates of Production Functions," *International Economic Review*, 8, pp. 568-598.

Ausiello, G., Marchetti-Spaccamela, A., and M. Protasi (1980). "Toward a Unified Approach for the Classification of NP-Complete Optimization Problems," *Theoretical Computer Science*, 12, pp. 83-96.

Bronars, S. (1987) "The Power of Nonparametric Tests of Preference Maximization," *Econometrica*, 55, pp. 693-698.

Caprara, A., Toth, P., and M. Fischetti (2000). "Algorithms for the Set Covering Problem," *Annals of Operations Research*, 98, pp. 352-371.

Caplin, A., and M. Dean (2008). "Information Search and the Choice Process," mimeo.

Choi, S., Gale, D., Fisman, R., and S. Kariv (2006). "Substantive and Procedural Rationality in Decisions under Uncertainty," mimeo.

Choi, S., Gale, D., Fisman, R., and S. Kariv (2007). "Consistency and Heterogeneity of Individual Behavior under Uncertainty," *American Economic Review*, 97(5), pp. 1921-1938.

Fisman, R., Kariv, S., and D. Markovits (2007) "Individual Preferences for Giving," *American Economic Review*, 97(2), pp. 153-158.

Garey, M.R. and D.S. Johnson. (1979). "Computers and Intractability: A Guide to the Theory of NP-Completeness," New York: W.H. Freeman.

Gross, J., and D. Kaiser (1996) "Two Simple Algorithms for Generating a Subset of Data Consistent with WARP and Other Binary Relations," *Journal of Business & Economic Statistics*, 14(2), pp. 251-255.

Houtman, M., and J. A. H. Maks (1985). "Determining all Maximal Data Subsets Consistent with Revealed Preference," *Kwantitatieve Methoden*, 19, pp. 89-104.

Johnson, D. (1975) "Finding All the Elementary Circuits of a Directed Graph," *SIAM Journal of Computing*, 4, pp. 77-84.

Varian, H. (1991) "Goodness-of-Fit for Revealed Preference Tests," University of Michigan CREST Working Paper # 13.